SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(138th Edition)

JUNE 2024

BY: CENTRAL LIBRARY

DELHI TECHNOLOGICAL UNIVERSITY
(FORMERLY *DELHI COLLEGE OF ENGINEERING)*
GOVT. OF N.C.T. OF DELHI
SHAHBAD DAULATPUR, MAIN BAWANA ROAD
DELHI 110042

PREFACE

This is the **One Hundred Thirty Eight** Issue of Current Awareness Bulletin started by Delhi Technological University, Central Library. The aim of the bulletin is to compile, preserve and disseminate information published by the faculty, students and alumni for mutual benefits. The bulletin also aims to propagate the intellectual contribution of Delhi Technological University (DTU) as a whole to the academia.

The bulletin contains information resources available in the internet in the form of articles, reports, presentations published in international journals, websites, etc. by the faculty and students of DTU. The publications of faculty and student which are not covered in this bulletin may be because of the reason that the full text either was not accessible or could not be searched by the search engine used by the library for this purpose.

The learned faculty and students are requested to provide their uncovered publications to the library either through email or in CD, etc. to make the bulletin more comprehensive.

This issue contains the information published during **June, 2023**. The arrangement of the contents is alphabetical. The full text of the article which is either subscribed by the university or available in the web is provided in this bulletin.

Central Library

CONTENTS

1. A Comparison of Transfer Learning Inspired CNN Architectures for 2D Image Workout Recognition, ***7.Aman Rehman***, ***3.Shailender Kumar*** and ***8.Lakshya Verma***, CSE, DTU

2. A Fractional-Order Meminductor Emulator with Applications in Chaotic Oscillator, Shireesh Kumar Rai, ***7.1.Bhawna Aggarwal*** and Rupam Das, Electronics, DTU

3. A Graded Index Hybrid Photonic Crystal Fiber for Supercontinuum Generation using AsSe2- As2S5, ***6.Drishti Singh Tomer*** and ***3.Ajeet Kumar***, A Physics, DTU

4. A lightweight convolutional swin transformer with cutmix augmentation and CBAM attention for compound emotion recognition, ***6.Nidhi*** and ***3.Bindu Verma***, IT, DTU

5. A Machine Learning and Augmented Reality based Framework for Multilingual Product Identification in Retail using Mobilenets and Vuforia, ***3.Geetanjali Bhola***, ***8.Amogh Bansal***, ***8.Divij Aggarwal*** and ***8.Gagan Kishor Upadhyay***, IT, DTU

6. A Metaheuristic Approach based Feasibility Assessment & Design of Solar, Wind and Grid Powered Charging of Electric Vehicles, Fayez Alanazi, ***6.1.Mohd Bilal***, Ammar Armghan and Md Reyaz Hussan, Electrical, DTU

7. A metasurface based close-proximity two-port circularly polarized MIMO antenna for mid-band sub-6 GHz 5G applications, ***6.Suresh Angadi***, ***3.Yashna Sharma***, ***3.N.S. Raghava*** and Thennarasan Sabapathy, ECE, DTU

8. A Novel Enhancement Mode GaN HEMT with MIS and p-GaN structure, ***6.Tanvika Garg*** and ***3.Sumit Kale***, Electronics, DTU

9. A novel nature-inspired nutcracker optimizer algorithm for congestion control in power system transmission lines, Vivek Kumar, R Narendra Rao, ***3.Ajendra Singh***, Vineet Shekher, Kaushik Paul, Pampa Sinha, Thamer AH Alghamdi and Almoataz Y Abdelaziz, Electrical, DTU

10. A novel software defect prediction model using two-phase grey wolf optimisation for feature selection, ***3.Ruchika Malhotra*** & ***6.Kishwar Khan***, Software Engineering, DTU

11. A Single Bitline Highly Stable, Low Power With High Speed Half-Select Disturb Free 11T SRAM Cell, ***6.LOKESH SONI*** and ***3.NEETA PANDEY***, Electronics, DTU

12. Adaptive control strategy for isolated renewable energy-based generation system with intermittency, Sombir Kundu, Ashutosh K. Giri, ***3.Madhusudan Singh*** and Sunil Kadiyan, Electrical, DTU

13. An Empirical Study to Enhance Leadership Qualities among Students by Improving the Social Intelligence Level of School Teachers: A Study of Select Schools in Delhi and NCR, Dr. Deepa Bisht and ***3.Dr. Meha Joshi***, DSM, DTU

14. Analysing Customer Satisfaction Towards Personal Loans: Evidence from Banking Industry, Divya, ***3.Deepti Aggrawal*** and Adarsh Anand, USME, DTU

15. Analysis of renewable energy and economic growth of Germany, ***6.Akanksha Singh*** and ***3.Nand Kumar***, Humanities, DTU

16. Analytical modeling of nucleation and growth of graphene layers on CNT array and its application in field emission of electrons, ***6.Shruti Sharma*** and ***3.Suresh C Sharma***, Applied Physics, DTU

17. As38 Se62 based segmented clad-graded index photonic crystal fber for supercontinuum generation covering 3–9.5 μm with moderate peak power, ***6.Akash Khamaru*** and ***3.Ajeet Kumar***, Applied Physics, DTU

18. BEACOMP: A Novel Textual Adversarial Attack Architecture for Unveiling the Fragility of Neural Text Classifiers, ***8.Rishabh Sharma***, ***8.Aatish Ghosh***, ***6.Ashish Bajaj*** and ***3.Dinesh Kumar Vishwakarma***, Electrical and IT, DTU

19. Bio Synergize: Microbial Synergy Driving Simultaneous Bioremediation and Nanoparticle Synthesis, ***7.Arif Khan***, ***7.Suman Yadav*** and ***3.Jai Gopal Sharma***, Biotechnology, DTU

20. Buildings of Distinct Plan Shapes Under Wind Load: Wind Interference Effects, Supriya Pal, ***3.Ritu Raj*** and , Alfia Bano, Civil, DTU

21. CDME-GAT: Context-Aware Depression Detection Using Multiembedding and Graph Attention Networks in Social Media Text, ***3.Minni Jain***, Siddh Jain, Amita Jain and Bhavuk Garg, CSE, DTU

22. Concentration of betanin from model beetroot extracts by using nanofiltration: parameter estimation and sensitivity analysis, ***6.Ashwani Kumar Tiwari*** and ***3.Manish Jain***, Applied Chemistry, DTU

23. Constraining Variable Generalized Chaplygin Gas model in Matter Creation Cosmology, ***7.Yogesh Bhardwaj*** and ***3.C. P. Singh***, Applied Mathematics, DTU

24. Control With Style: Style Embedding-based Variational Autoencoder for Controlled Stylized Caption Generation Framework, ***6.Dhruv Sharma***, ***3.Chhavi Dhiman*** and ***3.Dinesh Kumar***, Electronics, DTU

25. Cooperative MAC Protocol based on Best Data Rate (CMAC-DR), Nagesh R, M. N. Sree Rangaraju, ***6.N. R. Kushar Gowda*** and Vinayak Shekharappa Antin, DTU

26. Deep learning in neglected vector-borne diseases: a systematic review, ***8.Atmika Mishra***, ***8.Arya Pandey*** and ***3.Ruchika Malhotra***, Software Engineering, DTU

27. Deep learning-empowered intrusion detection framework for the Internet of Medical Things environment, ***3.Prashant Giridhar Shambharkar*** & ***6.Nikhil Sharma***, CSE, DTU

28. Design and analysis of a rectangular core refractive index-based PCF sensor for bio-sensing application, ***7.Neeraj Singh***, ***6.Akash Khamaru*** and ***3.Ajeet Kumar***, Applied Physics, DTU

29. Design of Dual-Band 4-Port Flexible MIMO Antenna for mm-Wave Technologies and Wearable Electronics, Rakesh N. Tiwari, Deepti Sharma, ***7.1.Prabhakar Singh*** and Pradeep Kumar, Electronics, DTU

30. Determination of Residual Stresses and Surface Roughness Using Chemical Assisted Ball End Magnetorheological Finishing, ***6.Anand Sharma*** and ***3.Mahendra Singh Niranjan***, Mechanical, DTU

31. Dy3+ doped KCa(PO3)3 phosphor for white light generation: structural and luminescent studies, ***7.Deepti***, ***7.Sandip Maurya***, ***6.Sheetal Kumari***, ***6.Pooja Rohilla***, Aman Prasad and ***3.A S Rao***, Applied Physics, DTU

32. Dynamic Resource-Aware Federated Framework for Secure and Sustainable Learning in Data Sensitive Applications, ***6.Manu Narula***, Jasraj Meena and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

33. Dynamics of Phase Transition in Quark-Gluon Plasma Droplet Formation under Magnetic Field, Agam K. Jha and ***8.Aviral Srivastava***, DTU

34. EFFECT OF AGRICULTURAL TRADE LIBERALIZATION ON FARM INCOME IN INDIA: AN EMPIRICAL ANALYSIS USING ARDL APPROACH, ***6.Hariom Prakash Verma*** and ***3.Nand Kumar***, Humanities, DTU

35. Evaluating the Environmental Impact of a Home Appliances: A Life Cycle Assessment with a New Green Product Design Framework, ***6.Mohd Tayyab***, ***3.Ranganath M Singari*** and Peer M Sathikh, Design, DTU

36. Examining the Dynamics of India's Major Exchange Rates Using Fourier Nonlinear Quantile Unit Root Test, Khyati Kathuria and ***3.Nand Kumar***, Humanities, DTU

37. Examining the structural viability of recycled fine aggregates in sustainable concrete, ***6.Harish Panghal*** and ***3.Awadhesh Kumar***, Civil, DTU

38. Exploring Biomedical Named Entity Recognition via SciSpaCy and BioBERT Models, Aman Jolly, Vikas Pandey, Indrasen Singh and ***6.Neha Sharma***, DTU

39. Fabrication of rGO-decorated hBNNS hybrid nanocomposite via organic–inorganic interfacial chemistry for enhanced electrocatalytic detection of carcinoembryonic antigen, ***6.Kanika Sharma***, ***3.Nitin K. Puri*** and ***3.Bharti Singh***, Applied Physics, DTU

40. Fe(III)-modified bamboo biochar for the removal of phosphate from synthetic and field stormwater runoff, ***6.Shivani Yadav***, ***7.Harsh Pipil***, ***3.A. K. Haritash*** and Krishna R. Reddy, Environmental, DTU

41. Fingerprint-Enabled ATM Network, ***8.Pragati Agrawal***, ***8.Rohan Saxena***, ***8.Saksham Agrawal*** and ***3.R.K. Singh***, Mechanical, DTU

42. FLOLSTM: Fuzzy logic-driven optimized LSTM for improved malicious traffic detection in hypervisor environments, ***6.Anumukonda Naga Seshu Kumar***, ***3.Rajesh Kumar Yadav*** and ***3.Nallanthighal Srinivasa Raghava***, CSE, Electronics, DTU

43. Harris Hawk Optimized CLAHE and Novel Score Level Fusion of Lightweight CNNs for Multimodal Biometric Recognition, ***3.Indu Singh***, ***8.Tanvi Singh***, ***8.Sunaina Luthra*** and ***8.Shruti Khatri***, CSE, DTU

44. IRIS: Intelligent Responsive Interface Solution, a promising Desktop Assistant for Differently-Abled, ***8.Aryan Karkra***, ***8.Raavi Singh*** & ***3.Dinesh K. Vishwakarma***, IT, DTU

45. Key Factors in Automobile Design: Expert Views on Enhancing User Experience during New Product Development, ***6.Shahid Ahmadm***, ***3.Ranganath M Singari*** and S.L. Bhandarkar, Design, DTU

46. Laboratory simulation to study the nonlinear evolution of right-hand circularly polarised wave in laser-produced plasmas, ***6.Aditya***, ***3.Suresh C Sharma***, Indraj Singh and R P Sharma, Applied Physics, DTU

47. LetsTalk: Understanding Social Media Usage of Self-Harm Users in India, ***3.Garima Chhikara*** and Abhijnan Chakraborty, DTU

48. Leveraging MediaPipe and YOLO Keypoint Detection in Ensemble Approaches for Workout Pose Recognition, ***7.Aman Rehman*** and ***3.Shailender Kumar***, CSE, DTU

49. Link chain driven reversible data hiding in encrypted images for high payload, ***6.Ankur***, ***3.Rajeev Kumar*** and Ajay K. Sharma, CSE, DTU

50. Machine learning: an advancement in biochemical engineering, ***8.Ritika Saha***, ***8.Ashutosh Chauhan*** and ***3.Smita Rastogi Verma***, Biotechnology, DTU

51. Magnesium oxide Nanoparticle synthesis using Rhizophora lamarckii plant extract and its characterization, ***8.Ankit Kundu*** & ***3.Jai Gopal Sharma***, Biotechnology, DTU

52. Meta-Learning-Based Physics-Informed Neural Network: Numerical Simulations of Initial Value Problems of Nonlinear Dynamical Systems Without Labeled Data and Correlation Analyses Worrawat Duanyai, Weon Keun Song, Thanadol Chitthamlerd and ***3.Girish Kumar***, Mechanical, DTU

53. Modified clustering and incentivized stable CH selection for reliable VANET communication, ***6.Ashish Kumar***, ***3.Shailender Kumar*** and Ram Shringar Raw, CSE, DTU

54. Modified-generative adversarial networks for imbalance text classification, ***6.Poonam Rani*** and ***3.Om Prakash Verma***, CSE and Electronics, DTU

55. Multimodal Architecture for Sentiment Recognition via employing Multiple Modalities, ***8.Samridhi Kapoor***, ***8.Shubham Gulati***, ***8.Sumit Verma***, ***8.Ananya Pandey*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

56. Navigating the Web of Disinformation and Misinformation: Large Language Models as Double-Edged Swords, ***8.SIDDHANT BIKRAM SHAH***, SURENDRABIKRAM THAPA, ASHISH ACHARYA, ***8.KRITESH RAUNIYAR***, SWETA POUDEL, SANDESH JAIN, ANUM MASOOD AND USMAN NASEEM, CSE, DTU

57. Numerical modeling of microstructure of slag incorporating cement paste, ***6.Andualem Yadeta***, ***3.Pradeep Goyal*** and ***3.Raju Sarkar***, Civil, DTU

58. Numerical modelling for triple hybrid gate optimization dielectric modulated junctionless gate all around SiNWFET based uricase and ChOX biosensor, Rishu Chaujar and ***6.Mekonnen Getnet Yirak***, Applied Physics, DTU

59. Numerical simulation of flow characteristics in nonprismatic compound channels, ***6.Vijay Kaushik***, ***3.Munendra Kumar*** and Bandita Naik, Civil, DTU

60. Optimization of battery/ultra-capacitor hybrid energy storage system for frequency response support in low-inertia microgrid, ***6.Philemon Yegon*** and ***3.Mukhtiar Singh***, Electrical, DTU

61. Performance analysis of LaFeO3 perovskite solar cells: A theoretical and experimental study, ***6.Rahul Kundara*** and ***3.Sarita Baghel***, Applied Physics, DTU

62. Performance Assessment of Underwater-to-Air Optical Wireless Communication System With the Effect of Solar Noise and Sea Surface Conditions, ***7.1.BHOGESWARA RAO ANGARA,*** PALANISAMY SHANMUGAM HARISHANKAR RAMACHANDRAN AND CHAVAPATI GOUSE SANDHANI, Electronics, DTU

63. Photonic Spin Hall Effect based AQI sensor: Effect of concentration of PM2.5, ***6.Lokesh Ahlawat***, ***3.Kamal Kishor***, Monu Nath Baitha and ***3.Ravindra Kumar Sinha***, Applied Physics, DTU

64. Physicochemical evaluation of polyvinyl alcohol films crosslinked with saturated and unsaturated dicarboxylic acids: A comparative study, ***6.Leela Gautam***, ***3.Sudhir G. Warkar*** and ***3.Manish Jain***, Applied Chemistry, DTU

65. Predicting Wealth Score from Remote Sensing Satellite Images and Household Survey Data Using Deep Learning, Shashank SHEKHAR, Pratibha SINGH, ***6.Rashmi MISHRA*** and Sunil KUMAR, CSE, DTU

66. Pressure-dependent bandgap characteristics in photonic crystals with sensing applications, ***6.Varnam Sherawat***, ***3.Renuka Bokolia*** and ***3.Ravindra Kumar Sinha***, Applied Physics, DTU

67. Probabilistic multiplicative unbalanced linguistic game using linguistic cloud model, ***6.Parul Chauhan*** and ***3.Anjana Gupta***, Applied Mathematics, DTU

68. Radially polarized femtosecond laser interaction with unmagnetized plasma slab and symmetric modes for enhanced terahertz field generation, ***6.Himank Sagar*** and ***3.Suresh C. Sharma***, Applied Physics, DTU

69. Role of polymers in encapsulation and shapestabilization of phase change materials (PCMs) for thermal energy storage applications: an overview, ***6.Surya Tanwar*** and ***3.Raminder Kaur***, Applied Chemistry, DTU

70. Shallowfake and deepfake image manipulation localization using noise and RGB-based dual branch method, ***6.Deepak Dagar*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

71. Short-term aggregate electric vehicle charging load forecasting in diverse conditions with minimal data using transfer and meta-learning, Shashank Narayana Gowda, ***8.Keshav Nath***, Chen Zhang, Rohan Gowda and Rajit Gadh, Mechanical, DTU

72. Synthesis of Graphene Quantum Dots (GQDs) from Paddy Straw for Bilirubin Detection, ***6.Isha***, ***6.Aneesha*** and ***3.Mohan Singh Mehata***, Applied Physics, DTU

73. Text mining based an automatic model for software vulnerability severity prediction, ***3.Ruchika Malhotra*** and ***7.Vidushi***, Software Engineering, DTU

74. The Influence Of Color On Visual Psychology And Cognitive Behavior: A Study In Paediatrics Hospital Environment, ***6.Monica Singh***, ***3.Ranganath M Singari*** and Mihir Bholey, Design, DTU

75. The rise of passive investing: a systematic literature review applying PRISMA framework, ***3.Priya Malhotra***, USME, DTU

76. Towards Optimal Resource Allocation: A Multi-agent DRL based Task Offloading Approach in Multi-UAV-Assisted MEC Networks, MUHAMMAD NAQQASH TARIQ, JINGYU WANG, SALMAN RAZA, ***7.1.MOHAMMAD SIRAJ***, MAJID ALTAMIMI and SAIFULLAH MEMON, Computer, DTU

77. Train Length and Service Frequency Optimization: Passenger Health Expense Perspective, ***7.Manish Kumar Sharma*** and Manvendra Singh, Electrical, DTU

78. Up-Conversion Luminescence in Sr9Y2W4O24:Ho3+, Yb3+ Phosphors for applications in Temperature Sensing and w-LEDs, ***6.Sheetal Kumari***, Aman Prasad, ***3.A. S. Rao*** and ***3.R. K. Sinha***, Applied Physics, DTU

79. Visual Sentiment Recognition via Popular Deep Models on the Memotion Dataset, ***6.Bhavana Verma***, ***3.Priyanka Meel*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

80. Weighted voting ensemble of hybrid CNN-LSTM Models for vision-based human activity recognition, ***8.Sajal Aggarwal***, ***3.Geetanjali Bhola*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

81. Workload dynamics implications in permissioned blockchain scalability and performance, ***6.Anita Thakur***, ***3.Virender Ranga*** and ***3.Ritu Agarwal***, IT, DTU

82. Zinc-Copper-Nickel Mixed Metal Oxide as Heterogeneous Catalytic Material for the Reductive Degradation of Nitroarene and Azo Dye, ***6.Jigyasa Pathak*** and ***3.Poonam Singh,*** Applied Chemistry, DTU

1. *Vice Chancellor*
2. *Pro Vice Chancellor*
3. *Faculty*
4. *Teaching-cum-Research Fellow*
5. *Asst. Librarian*
6. *Research Scholar*
7. *PG Scholar*
8. *Undergraduate Student*

1.1. *Ex Vice chancellor*
2.1. *Ex Pro Vice Chancellor*
3.1. *Ex Faculty*
4.1. *Alumni*
5.1 *Others*
6.1. *Ex Research Scholar*
7.1. *Ex PG Scholar*
8.1. *Ex Undergraduate Student*

2024 IEEE 9th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 5-7, 2024

A Comparison of Transfer Learning Inspired CNN Architectures for 2D Image Workout Recognition

Aman Rehman
Department of Computer Science and Engineering,
Delhi Technological University
Bawana Road, Delhi-110042, India
amanrehman18@gmail.com

Shailender Kumar
Professor, Department of Computer Science and Engineering,
Delhi Technological University
Bawana Road, Delhi-110042, India
shailenderkumar@dce.ac.in

Lakshya Verma
Department of Computer Science and Engineering,
Delhi Technological University
Bawana Road, Delhi-110042, India
lakshyaverma_co21a5_04@dtu.ac.in

2024 IEEE 9th International Conference for Convergence in Technology (I2CT) | 979-8-3503-9447-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/I2CT61223.2024.10543597

***Abstract*—Physical exercise is a key part of a healthy lifestyle, encompassing a wide array of activities from dance and aerobics, to weightlifting and sports. Proper form and posture are critical to ensure the safety of these exercises as well as making them effective. Accurate assessment of workout poses can provide invaluable feedback to individuals and fitness professionals alike. Performing exercises correctly also reduces the risk of injury, especially for people new to exercises. Additionally, fitness professionals can utilize this to remotely guide clients in virtual training sessions. Such systems can also act as a guide for people working out without any assistance. Deep learning approaches have been extensively used in computer vision as they show exceptional performance across a range of tasks. This may be used to automate the process of evaluating body postures during workouts. By leveraging these architectures, the aim is to develop an efficient and accurate system capable of precise pose identification. For this, we have taken a2D image dataset of 20 different workout postures. Weexperiment using 9 CNN architectures. EfficientNetB0 was found to have the greatest test accuracy of 91%.**

Keywords—deep learning, transfer learning, computer vision, pose detection

I. INTRODUCTION

Exercise routines are a vital part of staying in good physical health, reduce the risk of chronic illness, increase our energy, help us feel content and revitalize our body and mind. It takes proper form and skill to get the maximum benefit from any exercise.The same exercise may result in harm if carried out improperly. Especially when many individuals like to workout at home, or are traveling, or neither own a gym membership nor have access to a qualified trainer. The injuries caused due to improper form and guidance during exercises can do long-term damage. So, the attempt is to develop techniques that allow individuals to perform exercises safely. Here, we focus on realistic still images to develop machine learning models to do this task. There are video-based action recognition systems too, but in contrast, it may be more challenging for still images since they lack spatiotemporal information. There are several datasets which are used to perform exercise pose recognition from still images. However, we use a custom dataset as based on existing work, different papers have used different number of categories for different datasets and a wide variety of exercises. But, currently to our best knowledge, the number of categories we are including is higher than most related work. Also, there are inconsistencies in some and unavailability of some datasets. There is a need for a dataset that is less resource heavy, has different poses, people, lighting conditions, backgrounds and angles. Our dataset attempts to capture this wide variety.

Architecturally, implementing popular CNN architectures, specificallyVGG16, VGG19, ResNet50, DenseNet121, DenseNet201, MobileNetV2, InceptionV3, Xception and EfficientNetB0 via transfer learning are used to extract features from the images.Next, the models are adjusted to the subtleties of pose identification by training, hyperparameter tuning, and fine-tuning specific layers or the network as a whole.Finally, deploying the best model and continuously checking its performance allows possible improvement in accuracy and reliability in real-world applications.

II. RELATED WORK

Research on the identification of actions in still photos has regularly been conducted. Every study that pertains to the following makes use of still pictures. A residual neural network is used in Sreela et al. [1]'s image action recognition model to extract features efficiently. Succeeding in 15 Pascal VOC action courses, the model showed sensitivity to contextual changes and real-world imagery. By utilizing a residual neural network and a support vector machine classifier, the enhanced model achieved 66.1% accuracy in identifying human activities in a two-layer classification model.The research[2] uses ResNet as a feature extractor to identify human actions in still photos with accuracy of 61.07%. Activity recognition is augmented by approximating postures, and classification is handled by SVM. But these systems suffer in nuanced positions, truncations, and occlusions in complex situations.The two-stream convolutional neural network described in study[3] improves action recognition by using long-term fusion pooling to capture the temporal complexity of actions. Their method, called STDDCN, combines a knowledge distillation module with dense connectivity and multiscale information. Though it is successful in comprehending actions, it suffers from difficulties with computational load and large memory needs because of frame processing, even though it achieves great performance with fewer parameters. Their model yields an accuracy of 92.1%. Jose et al.[4] achieved 85% accuracy using the VGG-16 model to recognize 10 distinct yoga poses in still photos.

The study[5] used a neural network for sports picture classification across 6 categories using a pre-trained InceptionV3 for feature extraction. In sports picture categorization, their method yielded 96.64% accuracy rate.Podgorelec et al. [6]presented a CNN architecture based on the VGG-19 model with differential evolution method for classifying images. The CNN model's hyperparameters are optimized using the DE(differential evolution) approach after the training data has been fine-tuned using the TL method prior to classification. With 4 sports categories in their

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:31:17 UTC from IEEE Xplore. Restrictions apply.

unique dataset, they achieved 81.31% accuracy. There is room for improvement in accuracy.In order to identify workout actions, [7]suggests using a deep learning-based model called WorkoutNet. With 2796 photos representing 10 distinct exercise stances, the approach demonstrated an impressive validation accuracy of 92.75%. Using MediaPipe, Pham et al. [8] were able to extract skeletal sequences from RGB footage that was taken using a commercial camera. The DD-Net deep architecture is utilized for action recognition in the assessment section. The model's average accuracy is 98.33%.

Das et al.[9] developed a system that counts repetitions, analyzes comfort using heart rate data, and recognizes activities performed at home and at the gym. For indoor exercise recognition, they employed k-NN, SVM, and decision trees. They achieved remarkable accuracy of 99.4% for counting repetitions and 95.3% for activity detection. This technique is effective for usage in the gym or at home since it assists in selecting the proper weights for workouts.Using decision trees as the underlying classifier, Zhou et al.[10] used the confidence-based AdaBoost algorithm. They used inertial pressure sensors to detect motion during a variety of sports-related activities, with an emphasis on leg exercises. After six participants participated in twenty-four leg workouts, they achieved an accuracy of 81.7%. The utilization of sensor-based technologies in this technique shows potential for precise monitoring and analysis of leg exercises.The study's authors stressed the possibility of precisely measuring exercise time spent on routine tasks. They used Support Vector Machines (SVM) and Random Forest (RF) to reach an accuracy of about 96.2%. [11]speciallymade use of data gathered from wearable wristband sensors, demonstrating the potential to identify workout periods.

Nunavath et al. [12] have classified various physical activities using sensor data, especially accelerometer data. On the UCI-HAR dataset, they applied sensor-based methods such as Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN). Based on the gathered sensor data, their efforts yielded an accuracy of 84.89% in classifying these activities.Farhad et al.[13] used a pre-trained VGG-16 model to categorize a set of 18 distinct categories in a custom sports picture dataset. With this approach, they were able to classify these sports photos with 93% accuracy.A CNN-based technique for predicting human exercise postures shown in 2D photos was proposed by Haque et al. [14] used a carefully selected dataset with five distinct workout courses to achieve this. Their model properly identified these exercise postures with an accuracy of 82.68%.

III. Proposed Approach

The process used to complete the given task consists of multiple phases. To put it another way, as figure 1 illustrates, it includes gathering data, preparing it, choosing a model, training and honing it, assessing and then possibly deploying it.

A. Dataset Collection

We have taken variousexercise categories each containing 2D pose images which were pulled manually from Google search engine using Kaggle [15] as a reference. Thus, our dataset contains 1058 photos in total, divided into 20 classes, each with 55–60 images.As described above, the dataset attempts to capture a wide variety of exercises and different people performing them in various settings. The 20 exercise categories are as follows: Benchpress, bow, bridges, child's pose, cobra, crane pose, crunches, cycling, deadlift, donkey kicks, downdog, flutterkicks, lat-pulldowns, leg-extensions, plank, pullup, shoulder press, squat, standing toe reach andtricep-pushdown. The exercises are chosen to incorporate a variety of bodyweight and weight training exercises as well as stretches.

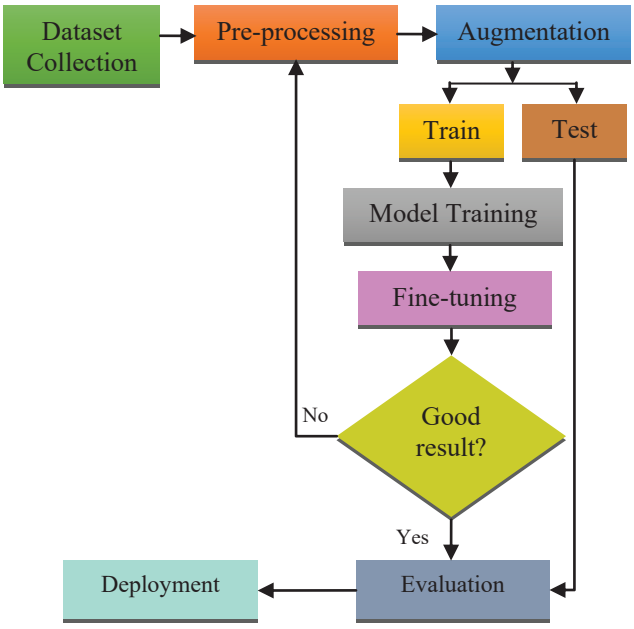


Fig. 1. Workflow

B. Data Processing and Augmentation

Images are loaded using Tensorflow and a standard size of 224x224 pixels is maintained. To avoid class-imbalanced splits, each image—labeled according to the exercise in the directory structure—is shuffled. The dataset split is 80:20 where 80% of the data is used for training, and the remaining 20% is used for testingafter normalizing to [-1, 1]. Before an image is fed into the models, it undergoes a predetermined set of augmentations, such as contrast and brightness fluctuation, random rotation, zooming, and flipping of the horizontal and vertical axes, to increase variability and facilitate generalization.

C. Model Architecture and Training

Multiple deep learning CNN architectures have been used to conduct a thorough study. ImageNet pre-trainedmodels were downloaded from Tensorflow. The advantage is that we can use the information already learnt by these models.To reduce spatial dimensions and focus attention on significant features, Global Average Pooling (GAP) layers were introduced.Also dropout layer is added to avoid overfitting.Later, dense layers are added with softmaxactivation function.The models are then trained and evaluated using accuracy metric.Categorical cross entropy loss function was used along with Adam optimizer.

D. Model Evaluation and Fine-Tuning

Model evaluation is performed on the test set. For a detailed analysis, we have used multiple methods on each model. We calculated the accuracy, the confusion matrices, class-wise accuracy and generated classification reports for each model.Training and testing accuracy and loss curves

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:31:17 UTC from IEEE Xplore. Restrictions apply.

were plotted to analyse the models for overfitting or underfitting. For the process of fine-tuning, we unfroze some of the pre-trained layers. After unfreezing, the parameters associated with those layers also become part of the training process. Multiple tests with different number of frozen layers were performed. After this, the models are further trained for more epochs. Varying improvement in accuracy was observed.

E. Comparative Study

A comparative study was done to obtain a more thorough understanding of the performance of each model best. Then, the best-performing model with the highest test accuracy was chosen, and its weights were saved for deployment later.

IV. Experimental Analysis

A. Model Architecture

For VGG16, the 16 layers are divided into 3 fully connected and 13 convolutional layers. It is often recommended for its simplicity and uniform structure. It uses small 3x3 convolutional filters throughout the network. For our task, we had a total of 14,724,948parameters and all of them were kept trainable.VGG19 is a larger variant of VGG16, with 3 additional convolutional layers. There are total of 20,034,644 parameters and all of them were kept trainable.

ResNet-50 introduces residual learningwith 50 layers, including GAP, fully connected, and 48 convolutional layers. Residual blocksuse a bottleneck design, consisting of a series of 1x1, 3x3, and 1x1 convolutions whichpreserves representational strength and decreases computational complexity.There are a total of 23,628,692 parameters, out of which 23,575,572 were kept trainable.

DenseNet contains densely connected layers, which ensure that each layer gets information directly from all layers before it. This promotes the reuse of features and eases information transfer across the network. For DenseNet121, there were a total of 7,633,108 parameters out of which 6,047,572 were kept trainable. For DenseNet201, there were a total of 20,995,604 parameters out of which 6,475,092 were kept trainable.

MobileNetV2 introduces depth-wise separable convolutions. Because of its reduced computation and parameter counts, it is ideal for mobile and edge devices and real-time applications with constrained processing capacity.There were a total of 2,283,604 parameters out of which 2,249,492were kept trainable.

InceptionV3 aims to capture multi-scale features efficiently using inception modules.They incorporate convolutions of varying filtersizes and concatenate feature maps. There were a total of 21,843,764parameters out of which21,809,332were kept trainable.

In Xception,depth wise separable convolutions are used. This minimizes the amount of parameters while capturing spatial interdependence efficiently and encouragingbetter information flow. The total number of parameters were20,902,460 out of which 20,847,932were kept trainable.

EfficientNet introduces the concept of compound scaling. This means that it scales the network's depth, resolution and width uniformly by a factor which can be learned. This guarantees a computationally efficient model which balances efficiency and complexity, improving performance under various resource restrictions. We have used EfficientNetB0 model whose architecture (figure2) consistently demonstrates good performance on benchmark datasets when it comes to image classification. We added our own GlobalAveragePooling2D layer, dropout layer (at 20% rate)and dense layers with softmax activation for20 exercisecategories. There were a total of 4,075,191 parameters out of which 4,033,168were kept trainable.

B. Requirements and Specification

Python version 3.10 was used to carry out the experiments. Hardware requirements include an Nvidia RTX 3080 GPU, a8-core Intel i7 processor, 64GB DDR4 RAM and Google ColabGPU for faster processing.

C. Result and Analysis

For all the 9 models, we have kept a learning rate of 0.001 with Adam optimizer while fine tuning with different number of layers and using categorical cross entropy loss.

We conducted an analysis of all the models used in our study for workout image recognition. Table 1's benchmark results demonstrate EfficientNetB0's superior performance, with the model attaining the maximum accuracy of 91% for top-1 classification. Among the best performers, DenseNet201 stood out with a 99.05% accuracy rate for top-5 classification closely followed by EfficicentNetB0 with 98.9%. We also present information on each model's precision, recall and F1 score.

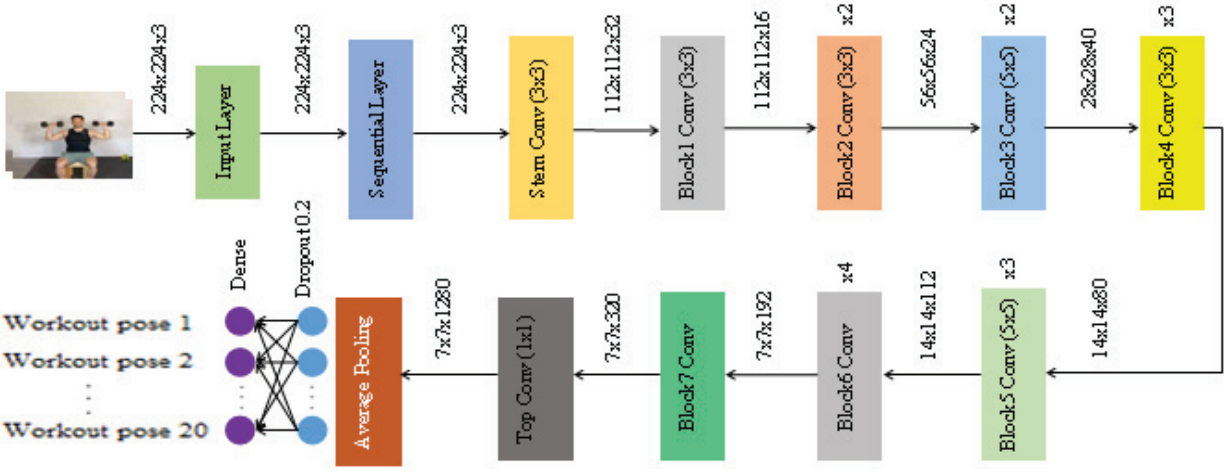


Fig. 2. EfficientNetB0 Model for our experiment

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:31:17 UTC from IEEE Xplore. Restrictions apply.

TABLE I. BENCHMARK RESULT OF 9 T-L MODELS

	Before fine-tuning	After fine-tuning				
Models	**Initial Accuracy**	**Top-1 Accuracy**	**Top-5 Accuracy**	**Precision**	**Recall**	**F1-Score**
VGG16	70.62%	79.62%	97.16%	0.88	0.81	0.81
VGG19	81.52%	83.89%	98.10%	0.85	0.84	0.84
ResNet50	76.30%	81.99%	97.16%	0.84	0.82	0.81
DenseNet121	66.35%	85.31%	98.00%	0.87	0.87	0.86
DenseNet201	72.51%	86.73%	**99.05%**	0.89	0.88	0.88
MobileNetV2	72.64%	76.42%	92.92%	0.81	0.76	0.75
InceptionV3	73.11%	84.43%	95.75%	0.85	0.84	0.83
Xception	76.30%	81.52%	96.68%	0.84	0.83	0.82
EfficientNetB0	86.73%	**91.00%**	98.90%	0.92	0.92	0.91

In the first training phase of EfficicentNetB0, the best model discovered, took place across 25 epochs, yielding an initial testing accuracy of 86.73% and a testing loss of 0.43. After then, the 150th layer was the focus of 10 more epochs of fine-tuning at a lower learning rate of 0.0001. With the use of checkpointing and optimal accuracy monitoring mechanisms, the procedure produced an astounding 99.65% fine-tuning training accuracy and 0.012 fine-tuning loss. Testing accuracy for the optimized model was 91.00%, and testing loss was 0.43, which was the same as during the first training phase. The model's accuracy was significantly increased overall through the fine-tuning procedure, indicating the success of the modifications made at this stage. The accuracy and loss curves in figures 3 and 4 graphically illustrate the model's development with fine-tuning; they show a clear tendency towards improved accuracy and lower loss.

In order to provide a more nuanced view of exercise-specific accuracy, we have included Table 2 that shows EfficientNetB0's accuracy for each workout class. The model performs well in a variety of exercises, with accuracy levels ranging from 69.23% to 100.00%. Most exercises yield greater than or equal to 80%, only tricep-pushdown shows a low performance of 69.23%. Examining the AUC-ROC curve as shown in figure 5 offers a thorough evaluation and shed light on the model's discriminative ability in this multi-class situation. The model's competency is demonstrated by notable precision, particularly the 100.00% accuracy for classes like Bow, Bridges, Child's Pose, Cobra, Crane Pose, Cycling, FlutterKicks and Pullup.

A breakdown of classifications is given by the confusion matrix(figure 6). Examining measures relevant to each class, such as precision and recall, we obtain detailed understanding of how well the model understands different types of exercises. Diagonal elements show the model correctly predicting poses. The off-diagonal elements point out areas that require improvement. It shows which incorrect class the model predicts for each exercise. The same follows for all the classes. This could be of interest as it may indicate similarity of movement or range of motion between exercises which may require more discriminating data to distinguish classes. This method provides a detailed insight of the strengths and limitations of the model.

TABLE II. POSE-WISE ACCURACY OF EFFICIENTNETB0 MODEL

Workout Pose	Accuracy
Benchpress	91.66%
Bow	100.00%
Bridges	100.00%
Child's Pose	100.00%
Cobra	100.00%
Crane Pose	100.00%
Crunches	90.00%
Cycling	100.00%
Deadlift	80.00%
Donkey kicks	90.00%
Downdog	84.61%
FlutterKicks	100.00%
Lat-pulldowns	88.88%
Leg-extensions	93.33%
Plank	82.35%
Pullup	100.00%
Shoulder Press	90.90%
Squat	90.00%
Standing toe reach	90.00%
Tricep-pushdown	69.23%

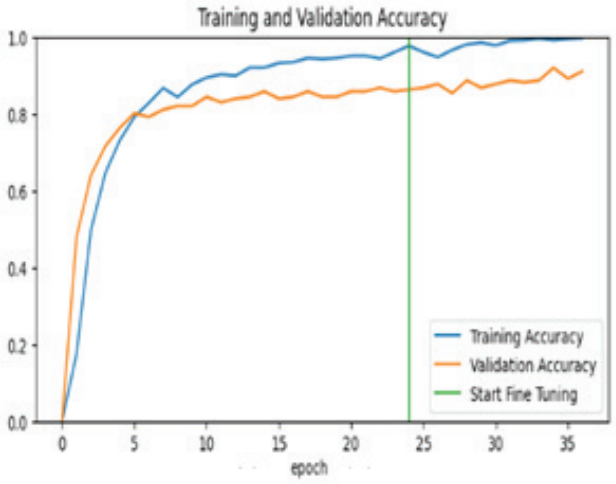


Fig. 3. Accuracy curve of EfficientNetB0 model

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:31:17 UTC from IEEE Xplore. Restrictions apply.

Fig. 4. Loss curve of EfficientNetB0 model

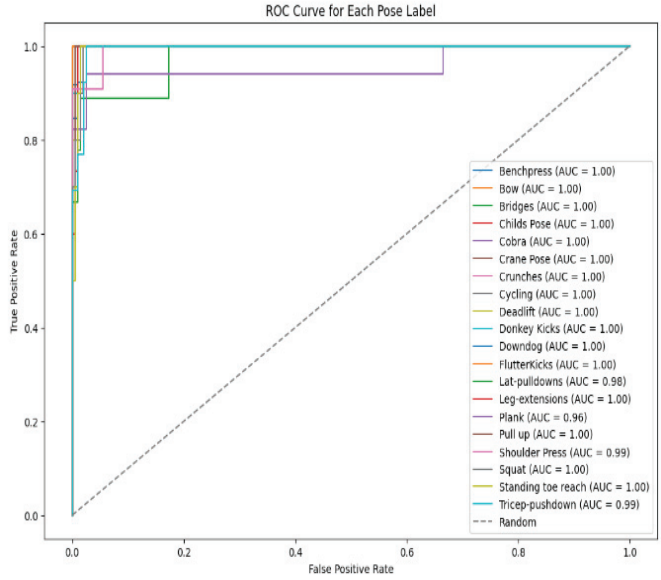


Fig. 5. AUC-ROC curve of EfficientNetB0 model

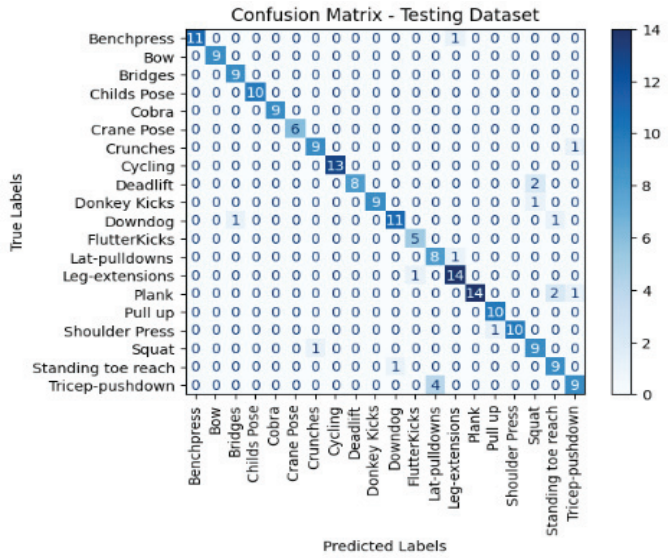


Fig. 6. Confusion-matrix for EfficientNetB0 model

D. Output

Figure 7 shows the true and predicted labels for images of different workout stances from the testing dataset. As demonstrated in the confusion matrix, there are incorrect predictions in the result as well. For classes where lower accuracy is obtained, it is useful to analyze possible similarities between images which may cause confusion.

Fig. 7. Actual vs Predicted pose of some test images

V. Conclusion

The presented experimental results reveal insights into the performance of prominent CNN architectures—VGG16, VGG19, ResNet50, DenseNet121, DenseNet201, MobileNetV2, InceptionV3, Xception, and EfficientNetB0 on workout pose detection. The dataset provides a challenging and diverse set of visual data for training and evaluation. The observed variations in test accuracies across models and epochs underscore the influence of architectural differences. EfficientNetB0 has demonstrated the strongest performance, achieving top 1 accuracy of 91%. Architectures like DenseNet121 and DenseNet201 notably performed worse than EfficientNetB0. These findings highlight the importance of selecting architectures that balances model complexity with the complexity of the problem, taking into account the diversity of exercises and the context of application.

A. Future work

Several avenues for future research exist. Firstly, the exploration of ensemble methods holds promise for further improving the overall accuracy. Combining predictions from multiple models can enhance generalization. Secondly, the integration of transformer-based models for vision, which is the latest development in computer vision tasks also holds promise. Specifically, vision transformers[14] have shown to beat previous state-of-the-art set by CNN architectures on multiple tasks. Furthermore, attention to data augmentation techniques could enhance model generalization. Specifcally, investigating results through an augmentation technique called RandAugment[13]. We may also attempt to improve model performance by incorporating keypoints. Finally, an exploration of explainability and interpretability may contribute to the transparency of the system and facilitate user trust and acceptance. Overall, the findings from this study lay the foundation for further advancements in workout

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:31:17 UTC from IEEE Xplore. Restrictions apply.

pose detection, encouraging the integration of novel architectures and methodologies to address the inherent challenges.

REFERENCES

[1] S. R. Sreela and S. M. Idicula, 'Action Recognition in Still Images using Residual Neural Network Features', *Procedia Comput. Sci.*, vol. 143, pp. 563–569, 2018, doi: 10.1016/j.procs.2018.10.432.

[2] W. Yang, Y. Wang, and G. Mori, 'Recognizing human actions from still images with latent poses', in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp. 2030–2037. doi: 10.1109/CVPR.2010.5539879.

[3] H. Kwon, Y. Kim, J. S. Lee, and M. Cho, 'First Person Action Recognition via Two-stream ConvNet with Long-term Fusion Pooling', *Pattern Recognit. Lett.*, vol. 112, pp. 161–167, Sep. 2018, doi: 10.1016/j.patrec.2018.07.011.

[4] J. Jose and S. Shailesh, 'Yoga Asana Identification: A Deep Learning Approach', *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1110, no. 1, p. 012002, Mar. 2021, doi: 10.1088/1757-899X/1110/1/012002.

[5] K. Joshi, V. Tripathi, C. Bose, and C. Bhardwaj, 'Robust Sports Image Classification Using InceptionV3 and Neural Networks', *Procedia Comput. Sci.*, vol. 167, pp. 2374–2381, 2020, doi: 10.1016/j.procs.2020.03.290.

[6] V. Podgorelec, Š. Pečnik, and G. Vrbančič, 'Classification of Similar Sports Images Using Convolutional Neural Network with Hyper-Parameter Optimization', *Appl. Sci.*, vol. 10, no. 23, Art. no. 23, Jan. 2020, doi: 10.3390/app10238494.

[7] A. Dey, A. Dutta, and S. Biswas, 'WorkoutNet: A Deep Learning Model for the Recognition of Workout Actions from Still Images', in *2023 3rd International Conference on Intelligent Technologies (CONIT)*, Jun. 2023, pp. 1–8. doi: 10.1109/CONIT59222.2023.10205926.

[8] Q.-T. Pham *et al.*, 'Automatic recognition and assessment of physical exercises from RGB images', in *2022 IEEE Ninth International Conference on Communications and Electronics (ICCE)*, Jul. 2022, pp. 349–354. doi: 10.1109/ICCE55644.2022.9852094.

[9] D. Das, S. M. Busetty, V. Bharti, and P. K. Hegde, 'Strength Training: A Fitness Application for Indoor Based Exercise Recognition and Comfort Analysis', in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2017, pp. 1126–1129. doi: 10.1109/ICMLA.2017.00012.

[10] B. Zhou, M. Sundholm, J. Cheng, H. Cruz, and P. Lukowicz, 'Never skip leg day: A novel wearable approach to monitoring gym leg exercises', in *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Mar. 2016, pp. 1–9. doi: 10.1109/PERCOM.2016.7456520.

[11] Y. Yoshida and E. Yuda, 'Workout Detection by Wearable Device Data Using Machine Learning', *Appl. Sci.*, vol. 13, no. 7, Art. no. 7, Jan. 2023, doi: 10.3390/app13074280.

[12] V. Nunavath *et al.*, 'Deep Learning for Classifying Physical Activities from Accelerometer Data', *Sensors*, vol. 21, no. 16, p. 5564, Aug. 2021, doi: 10.3390/s21165564.

[13] M. Yasir Farhad, S. Hossain, MD. Rezaul Karim Tanvir, and S. Ameen Chowdhury, 'Sports-Net18: Various Sports Classification using Transfer Learning', in *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*, Dec. 2020, pp. 1–4. doi: 10.1109/STI50764.2020.9350415.

[14] S. Haque, A. S. A. Rabby, M. A. Laboni, N. Neehal, and S. A. Hossain, 'ExNET: Deep Neural Network for Exercise Pose Detection', in *Recent Trends in Image Processing and Pattern Recognition*, K. C. Santosh and R. S. Hegadi, Eds., in Communications in Computer and Information Science. Singapore: Springer, 2019, pp. 186–193. doi: 10.1007/978-981-13-9181-1_17.

[15] 'Workout/Exercise Images'. Accessed: Dec. 19, 2023. [Online]. Available: https://www.kaggle.com/datasets/hasyimabdillah/workoutexercises-images

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:31:17 UTC from IEEE Xplore. Restrictions apply.

A Fractional-Order Meminductor Emulator with Applications in Chaotic Oscillator

Shireesh Kumar Rai[a], Bhawna Aggarwal[b*], Rupam Das[c]

[a]Electronics and Communication Engineering Department, Thapar Institute of Engineering and Technology, Punjab, India

[b]Electronics and Communication Engineering Department, Netaji Subhas University of Technology, Delhi, India

[c]Electronics and Communication Engineering Department, Asansol Engineering College, West Bengal, India

Abstract: This paper introduces a fractional-order meminductor emulator (FOMI) with both grounded and floating configurations, showcasing its applications in a chaotic oscillator. It employs one voltage differencing inverted buffered amplifier, a current follower, a fractional-capacitor, and a conventional capacitor. The transient analysis, pinched hysteresis loops, and non-volatile characteristics obtained for the suggested FOMI serve as clear indicators of the circuit's effective operation. The variations in pinched hysteresis loops with changes in the fractional-order (α) and frequency also support the theoretical concepts. The circuit's robust performance has been assessed by analyzing simulation results under varying conditions of temperature and supply voltage. The analysis also involves examining the impact of varying capacitor values on pinched hysteresis loops. Monte Carlo and corner analyses support the robust behaviour of the circuit. The suggested FOMI's potential use has been demonstrated using the chaotic oscillator circuit.

Keywords: Pinched hysteresis loop, chaotic oscillator, memristor, meminductor, memcapacitor, fractional-order.

1. Introduction

The exploration of novel circuit components and architectures is a continuous pursuit in the field of electronic design, leading to advancements with applications across diverse disciplines. While traditional approaches focus on integer-order differentiation and integration, the potential of non-integer, or fractional-order, calculus has gained significant interest due to its ability to model real-world phenomena more accurately. This branch of mathematics, known as fractional calculus, offers greater flexibility by introducing the concept of fractional-order derivatives and integrals. Its impact has transcended boundaries, finding applications in various scientific and engineering fields, including material science, biology, and control theory [1].

a. Tel.: +919911941540, E-mail addresses: skumar.rai@thapar.edu,

b. Corresponding author. Tel.: +9180106269659, E-mail addresses: bhawna.aggarwal@nsut.ac.in,

c. Tel.: +919064120428, E-mail addresses: das_rupam@rediffmail.com

To take the advantages offered by fractional calculus, in circuit theory, fractional-order capacitors (FOCs) have been realized. These elements, denoted as C_α where α lies between 0 and 1, are also known as constant phase elements (CPEs). Their impedance characteristic, $1/s^\alpha C_\alpha$, introduces a constant phase shift of $-\alpha\pi/2$. However, due to the absence of readily available physical components exhibiting fractional-order behaviour, various techniques based on RC networks have been developed to approximate FOCs through continuous fraction expansion [2, 3]. Notably, Foster and Cauer networks utilize resistors and capacitors to achieve this approximation, with a trade-off between network complexity and accuracy. While higher-order networks provide better approximations, they also require more components. The foster form of fifth order is often realized with the help of six resistors and five capacitors.

The potential of fractional-order elements extends beyond capacitors, finding application in the recently discovered memristor. Unlike traditional capacitors and inductors that store energy, memristors possess the unique ability to store information, making them highly attractive for low-power neuromorphic computing applications. Their capability to retain and process information within a compact footprint further enhances their appeal. While conventional circuit components lack the captivating features of memristors, these were first envisioned by Prof. Leon Chua in 1971 [4] and eventually physically realized using Titanium dioxide (TiO2) by HP labs after a 37-year gap [5]. Today, they are readily available as commercial components. However, memristors' counterparts, memcapacitors and meminductors, are not yet widely available, prompting researchers and engineers to develop emulator circuits that replicate their expected characteristics. While numerous integer-order memcapacitor and meminductor emulator designs exist, research on their fractional-order counterparts remains scarce.

The motivation for this paper stems from this noticeable gap in research concerning fractional-order meminductors. In addition to the fact that fractional-order meminductors remain underexplored as compared to their integer-order counterpart, it has been analyzed that the existing fractional-order meminductor designs are often intricate, limiting their frequency responses to the kHz range. *This paper addresses this critical gap by introducing a simplified circuit for a FOMI with a superior frequency response.* Unlike existing fractional-order meminductors, the proposed design directs for a more straightforward architecture, ensuring broader applicability across broad frequency ranges. Beyond overcoming the

limitations of existing designs, the untapped potential of fractional order meminductors has been explored compared to their integer-order counterparts. Specifically, in chaotic circuits, fractional-order meminductors hold immense promise for generating complex dynamics and offering greater flexibility in controlling chaotic behavior. To achieve these objectives, a new design of meminductor emulator has been realized that leverages two well-established active building blocks: a voltage differencing inverted buffered amplifier (VDIBA) and a current follower (CF). Additionally, the design incorporates both a fractional-order capacitor and a conventional capacitor. The resulting circuit demonstrates robust performance, as verified through simulations conducted under varying temperature and supply voltage conditions. Finally, to showcase the versatility of the proposed fractional-order meminductor emulator (FOMI), its potential application in chaotic oscillator has been demonstrated.

A comparison summary of the suggested FOMI with existing ones are presented in Table 1. The observations of Table 1 are provided below:

1. For fractional meminductor emulators described in the literature [6-13], the greatest frequency for PHLs is restricted up to kHz only. In comparison, the suggested FOMI operates till 7MHz.
2. The suggested circuit is free from the requirement of memristors. This is in contrast to the known meminductor emulators [6], [7], and [12] that require memristors for their proper functioning.
3. The meminductor emulators [6], [11], and [13]) involve multipliers, active building blocks, along with a lot of passive components, nonetheless, the proposed FOMI has a fairly simple architecture and only requires a few components.

The content of the paper is arranged into seven sections which start from the first section of the introduction. The short review is given in Section 2 for the ease of readers. Section 3 presents the traits of active building blocks (ABBs) and their CMOS realizations. Section 4 discusses the operational concept and mathematical formulations of the proposed FOMI. Section 5 presents the simulation results. Applications of the proposed floating meminductor emulator namely chaotic oscillator is given in Section 6. Section 7 contains the paper's conclusion.

2. Short review of FOMI

Fractional meminductors extend the concept of traditional inductors and capacitors by incorporating fractional calculus principles. The application of fractional calculus principles

enables the analysis of time-domain behaviour and fractional-order dynamics, thereby enhancing modelling capabilities for these devices and offering potential solutions across a range of applications. In fractional calculus, the fractional derivative of a function is expressed in terms of the differentiation of non-integer order. Similarly, the fractional integral is an integration of non-integer order. FOMI is a device that exhibits memory effects in its current-flux relationship, similar to a traditional inductor, but with the inclusion of fractional-order derivatives. The FOMI's fractional nature enables it to capture more intricate and non-linear system behaviours. The FOMI can be tuned to have particular memory characteristics and frequency responses, allowing for more flexibility in system design and analysis.

The fractional derivative of a function *f(t)* with respect to time *t* of order α, denoted by D_t^{α}, is defined using the Caputo or Riemann-Liouville fractional derivative operators. The Caputo fractional derivative operator is commonly used and is defined as follows:

$$D_t^{\alpha} f(t) = \frac{1}{\Gamma(n-\alpha)} \int_0^t \frac{f^{(n)}(\tau)}{(t-\tau)^{\alpha+1-n}} d\tau \quad (1)$$

where, $\Gamma(.)$ is gamma function and $n\text{-}1 \leq \alpha \leq n$ is the fractional order.

In the case of fractional meminductors, the fractional derivative term is incorporated into the constitutive relation between the current (I) and flux (φ) of the meminductor. The fractional-order derivative of current (I) with respect to time (t) is related to the fractional-order derivative of voltage (V) by a constant fractional-order meminductance (L^{α}) [14]:

$$I(t) = L^{-\alpha} \cdot D^{\alpha-1} V(t) \quad (2a)$$

here, $D^{\alpha-1}$ represents fractional derivative. Considering the fact that flux is time integral of voltage, Equation (2a) can be re-written as:

$$I(t) = L^{-\alpha} \cdot D^{\alpha} \varphi(t) \quad (2b)$$

The fractional order α, which controls the memory characteristics, frequency response, and dynamic behaviour of the device, affects the distinct characteristics and behaviour of fractional-order meminductors. The desired characteristics of meminductors can be attained by choosing suitable values for α.

3. Overview of ABBs

VDIBA and CF have been used in the designing of the proposed FOMI.

3.1 VDIBA

The VDIBA is a four-terminal amplifier, as illustrated in Fig. 1 (a). The positive and negative high-impedance input voltage terminals are denoted as V+ and V-, respectively. The output voltage terminal W- has a low impedance, while the intermediate terminal 'Z' exhibits high impedance. In terms of implementation, its input stage comprises an OTA with a differential input and a single output. This OTA converts the differential signal into a corresponding current at the 'Z' terminal. The voltage at the W- terminal of the output stage unity-gain inverting voltage buffer is the inverse of the voltage observed at the 'Z' terminal. An additional biasing terminal V_B can be used to control the transconductance parameter (G_m). Applications requiring resistor-less and electronically regulated circuits encourage the use of VDIBA because its G_m may be electronically controlled by a biasing voltage.

The equations defining the port characteristics of VDIBA are given as:

$$I_{V+} = I_{V-} = 0 \tag{3}$$

$$I_Z = G_m\left(V_{V+} - V_{V-}\right) \tag{4}$$

$$V_{W-} = -\beta V_Z \tag{5}$$

$$G_m = \frac{K}{2\sqrt{2}}\left(V_B - V_{SS} - 2V_{th}\right) \tag{6}$$

where K is a technological parameter and its value is given by $\mu_n \cdot C_{ox} \cdot W/L$, with μ_n representing the mobility of charge carriers, C_{ox} is per unit area capacitance due to the oxide layer, W is the width and L is the channel length of the symmetric MOSFETs constituting differential pair of OTA. V_{SS} is the negative supply voltage used for biasing of VDIBA and V_{th} is the threshold voltage of the MOSFETs that govern the gain of the amplifier. β denotes the gain of 'W-' terminal in the non-ideal case and is ideally unity. The CMOS circuit of VDIBA depicting all its port terminals is shown in Fig. 1 (b).

3.2 CF Block

The current at the output terminal follows the input current in a traditional CF, which has a low input impedance and a very high output impedance. The number of output terminals and their current polarity can be adjusted by using different combinations of current mirrors. The block diagram and CMOS circuit of a four-terminal CF are shown in Figs. 2 (a) and (b), respectively.

Here ‘X’ is a low-impedance input current port, Z_{1+} and Z_{2+} are two high-impedance output ports with positive current polarity, and Z- is a high-impedance output port with negative current polarity. The port equations describing the current at various ports of CF are given as:

$$V_X = 0 \tag{7}$$

$$I_{Z1+} = I_{Z2+} = \gamma \cdot I_X \tag{8}$$

$$I_{Z-} = -\gamma \cdot I_X \tag{9}$$

here, γ is the current gain of non-ideal CF and has a value equal to unity in case of the ideal output.

4. Proposed FOMI

Electronic components called fractional-order meminductors have fractional-order dynamics in terms of memory effects and nonlinearity. Fractional-order dynamics can offer greater flexibility and control for simulating memory effects in the context of meminductor emulators. The fractional meminductor emulator can simulate a variety of meminductive behaviours, including varied degrees of memory and nonlinearity, by varying the order of the fractional calculus operator. A few designs [6-13] are available for floating-order meminductors. It is well-known that fractional-order meminductor provides many advantages over integer-order meminductors such as better control over pinched hysteresis loops, complex chaos generation in chaotic oscillators for secure communication, and additional variable (α) available for controlling many circuit parameters, etc. Therefore, in this paper, a FOMI has been proposed. This meminductor is designed using VDIBA and CF as active blocks. Along with active blocks, the proposed emulator employs a fractional capacitor and a conventional capacitor. The fractional capacitor (C_α) is connected to VDIBA's ‘Z’ terminal to produce the fractional behaviour. The VDIBA is used in view of charging C_α and its internal buffer helps in copying the voltage generated at the ‘Z’ terminal to the ‘W-’ terminal. The proportional voltage produces a current I_X for the CF that is utilized to charge the capacitor C_2. The capacitor C_2 is used to remember the state of the circuit. The voltage across C_2 is further utilized to control the fractional meminductance of the proposed FOMI.

4.1 Proposed grounded FOMI

The complete circuit of the suggested grounded FOMI designed using VDIBA and CF has been presented in Fig. 3(a). The depiction of the Foster-I realization employed for the

implementation of C_α is illustrated in Fig. 3(b). The order of the RC network shows a significant compromise between the accuracy of the achieved fractional capacitor and the intricacy of the design. In the presented implementation, a 5th order RC network has been selected for the realization of the fractional capacitor. This choice has been governed by the fact that a 5th order Foster-I RC network offers a precise emulation of fractional capacitors across a broad frequency spectrum.

In the proposed meminductor of Fig. 3(a), the incremental configuration is obtained by connecting the V- terminal to the input and the V+ terminal to the ground (ⓐ → ⓓ and ⓑ → ⓒ). In order to obtain decremental configuration, the input is connected to V+ while V- terminal is grounded (ⓐ → ⓒ and ⓑ → ⓓ). The fractional capacitor C_α is charged by the voltage reflected at the 'Z' terminal. The 'X' terminal of the CF draws current from 'W-' and is reflected to 'Z_{1+}', 'Z_{2+}', and 'Z-' terminals. The capacitor C_2 is charged by the current drawn from 'W-' terminal. The voltage developed across C_2 biases the VDIBA and controls the transconductance G_m as per Equation (6). The feedback mechanism created by the charge stored at C_2 provides the basis for memory in this meminductor.

To verify the working of the circuit, a detailed mathematical analysis of the incremental meminductor is carried out as follows:

Using Equation (4) and analyzing from Fig. 3(a) that $V_{V-} = V_{in}$ and $V_{V+} = 0$, the expression obtained is:

$$I_Z = G_m\left(-V_{in}\right) \tag{10}$$

Using Equation (5), the voltage across C_α can be written in s-domain as:

$$V_Z\left(s\right) = \frac{1}{s^\alpha C_\alpha}\cdot I_Z = -\frac{G_m V_{in}}{s^\alpha C_\alpha} \tag{11}$$

where α is the fractional order of C_α.

Since, the flux at the input is represented as: $\phi_{in}(s) = \frac{V_{in}}{s}$, Equation (11) gets modified as:

$$V_Z\left(s\right) = \frac{-G_m}{s^{\alpha-1} C_\alpha}\cdot \varphi_{in}\left(s\right) \tag{12}$$

Direct analysis of Fig. 3(a) yields the current drawn by CF as:

$$I_X = \frac{V_{W-}}{R_1} \tag{13}$$

where R_1 is the parasitic resistance seen between 'W-' and 'X' terminals and is given as:

$$R_1 = R_{W-} + R_X I_X \tag{14}$$

here, R_{W-} and R_X denote the parasitic resistances at 'W-' and 'X' terminals, respectively. The voltage across the capacitor C_2 ($V_{C2}(s)$) is given by:

$$V_{C2}(s) = V_B(s) = \frac{1}{s \cdot C_2} \cdot I_{Z2+} \tag{15}$$

Using Equation (8) and Equation (13), Equation (15) can be rewritten as:

$$V_B(s) = \frac{\gamma}{s \cdot C_2} \cdot \frac{1}{R_1} \cdot V_{W-} \tag{16}$$

Substituting V_{W-} from port Equation (5) and V_Z from Equation (12), Equation (16) can be expressed as:

$$V_B(s) = \frac{\gamma \cdot \beta \cdot G_m}{s^{\alpha} R_1 \cdot C_{\alpha} \cdot C_2} \cdot \varphi_{in}(s) \tag{17}$$

Since, $\frac{\phi_{in}}{s} = \rho(s)$, where $\rho(s)$ is the time domain integral of flux, Equation (17) can be redefined as:

$$V_B(s) = \frac{\gamma \cdot \beta \cdot G_m}{s^{\alpha-1} R_1 \cdot C_{\alpha} \cdot C_2} \cdot \rho(s) \tag{18}$$

The transconductance parameter (G_m) of VDIBA is dependent on biasing voltage V_B, as seen in Equation (6). Substituting value of V_B from Equation (18) in Equation (6) yields:

$$G_m(s) = \frac{K}{2\sqrt{2}} \left(\frac{\gamma \cdot \beta \cdot G_m}{s^{\alpha-1} R_1 \cdot C_{\alpha} \cdot C_2} \cdot \rho(s) - V_{SS} - 2V_{th} \right) \tag{19}$$

From Fig. 3(a), $I_{in} = -I_{Z+}$. Further combining Equations (5), (9), (12) and (13), we get:

$$I_{in}(s) = \frac{\gamma \cdot \beta \cdot G_m \cdot \varphi_{in}(s)}{s^{\alpha-1} R_1 \cdot C_{\alpha}} \tag{20}$$

Substituting the value of G_m from Equation (19) in Equation (20), yields:

$$I_{in}(s) = \frac{\gamma \cdot \beta}{s^{\alpha-1} R_1 \cdot C_{\alpha}} \frac{K}{2\sqrt{2}} \left(\frac{\gamma \cdot \beta \cdot G_m}{s^{\alpha-1} R_1 \cdot C_{\alpha} \cdot C_2} \cdot \rho(s) - V_{SS} - 2V_{th} \right) \cdot \varphi_{in}(s) \tag{21}$$

The relation between current and induced flux in meminductor in terms of meminductance (L_M) is given as:

$$I(s) = L_M^{-1} \cdot \varphi(s) \tag{22}$$

Considering Equations (21) and (22), the inverse meminductance of the proposed meminductor is expressed as:

$$L_M^{-1} = \frac{\gamma \cdot \beta}{s^{\alpha-1} R_1 \cdot C_\alpha} \frac{K}{2\sqrt{2}} \Bigg(\underbrace{V_{SS} + 2V_{th}}_{\text{fixed part}} - \underbrace{\frac{\gamma \cdot \beta \cdot G_m}{s^{\alpha-1} R_1 \cdot C_\alpha \cdot C_2} \cdot \rho(s)}_{\text{variable part}} \Bigg) \tag{23}$$

where α is fractional order, β is the voltage gain of VDIBA and γ is the current gain of CF. Equation (23) shows that the meminductance of the proposed meminductor can be controlled by fixed parameters (supply voltage and threshold voltage) and variable parameters (input flux). The fractional power (α) helps in precise control of the meminductance. The circuit shown in Fig. 3 (a) can be used in decremental mode by connecting ⓐ → ⓒ and ⓑ → ⓓ. A similar mathematical analysis as carried out for incremental configuration, yields meminductance for decremental configuration as:

$$L_M^{-1} = \frac{\gamma \cdot \beta}{s^{\alpha-1} R_1 \cdot C_\alpha} \frac{K}{2\sqrt{2}} \Bigg(\underbrace{V_{SS} + 2V_{th}}_{\text{fixed part}} + \underbrace{\frac{\gamma \cdot \beta \cdot G_m}{s^{\alpha-1} R_1 \cdot C_\alpha \cdot C_2} \cdot \rho(s)}_{\text{variable part}} \Bigg) \tag{24}$$

From Equation (24) it can be observed that both fixed and variable components are negative, leading to decremental behaviour of the proposed fractional meminductor.

4.2 Proposed floating FOMI

The suggested floating type FOMI is shown in Fig. 4. When the circuit of Fig. 4 is subjected to the same mathematical analysis that was done for the grounded type, the relationship obtained between flux and current for incremental configuration is provided below:

$$L_M^{-1} = \frac{\gamma \cdot \beta}{s^{\alpha-1} R_1 \cdot C_\alpha} \frac{K}{2\sqrt{2}} \Bigg(\underbrace{V_{SS} + 2V_{th}}_{\text{fixed part}} + \underbrace{\frac{\gamma \cdot \beta \cdot G_m}{s^{\alpha-1} R_1 \cdot C_\alpha \cdot C_2} \cdot \rho(s)}_{\text{variable part}} \Bigg) \tag{25}$$

By flipping the switch as indicated in the figure, the proposed emulator shows decremental behaviour and its meminductance can be expressed as:

$$L_M^{-1} = -\frac{\gamma \cdot \beta}{s^{\alpha-1} R_1 \cdot C_\alpha} \frac{K}{2\sqrt{2}} \Bigg(\underbrace{V_{SS} + 2V_{th}}_{\text{fixed part}} + \underbrace{\frac{\gamma \cdot \beta \cdot G_m}{s^{\alpha-1} R_1 \cdot C_\alpha \cdot C_2} \cdot \rho(s)}_{\text{variable part}} \Bigg) \tag{26}$$

In the proposed FOMI, Equations (23) to (26) show that the fractional s-domain factor of $s^{\alpha-1}$ offers greater flexibility in modelling complex systems. This aids in capturing a broader range of system behaviours that integer-order models are unable to faithfully capture. Additionally, it offers greater accuracy and increased adaptability in simulating real-world

systems, particularly those displaying non-linear and non-integer behaviours. However, fabricating such emulators with precise fractional-order characteristics is a complex task, leading to increase in manufacturing costs. Additionally, fractional-order circuits are generally more sensitive to variations in component values compared to their integer-order counterparts. Even minor discrepancies in resistance, capacitance, or other parameters can significantly affect the behaviour of the emulator, posing challenges in practical implementations. Furthermore, commercially available components specifically designed for fractional-order circuits are limited compared to those readily available for integer-order circuits. Furthermore, existing circuit design methodologies, simulation tools, and standard components are primarily tailored for integer-order circuits making the analysis of these circuits a challenging task.

5. Simulation results

Simulations using the LTspice tool have been performed to test the operation of the proposed fractional-order meminductor circuit. The blocks VDIBA, and CF have been simulated with supply rails of ±0.9V utilizing TSMC's 180nm technology parameters. Table 2 displays the size of the MOSFETs utilized in the active blocks. The capacitance value C_2 is set to 10 pF in order to achieve pinched hysteresis lobes. A fifth-order Foster-I form is used to implement C_α linked to the 'Z' terminal of VDIBA, as displayed in Fig. 3(b). The values of R and C for the Foster-I form have been calculated using MATLAB [15]. These values for C_α = 10pF, for different values of α have been presented in Table 3. Unless otherwise stated, all the simulated results have been plotted with α = 0.5 for the sake of simplicity.

5.1 Simulation results of proposed floating FOMI in incremental configuration

The curves representing transient analysis of the proposed floating FOMI for α = 0.5, plotted with a 100kHz sinusoidal signal have been displayed in Fig. 5. This analysis reveals intricate details of the proposed floating FOMI's temporal behaviour showcasing the unique relationship between voltage, flux, and current waveforms. In this figure it can be observed that the voltage waveform leads the flux and current waveforms. This is a characteristic that is observed in inductive circuits. The non-sinusoidal nature of the current waveform is attributed to the dynamic fluctuations in the inductance value over time, providing a comprehensive understanding of the circuit's transient response. To check the non-volatile behaviour of the proposed emulator, a step signal with 15µs period and 1µs ON time is

applied at the input terminal and the observed response is shown in Fig. 6. The figure illustrates a key aspect of the emulator's behavior. Specifically, during the ON time of the step signal, the emulator's meminductance value increases. Meminductance is a property similar to inductance but with memory effects, implying that the device "remembers" its previous states. What is particularly noteworthy is that when the pulse train is turned OFF, the emulator retains its meminductance value from the previous ON state. This persistence of the meminductive property during the OFF state demonstrates the non-volatile behaviour of the proposed emulator. This ensures that the emulator has the ability to "remember" and maintain its meminductive state even after the input signal is no longer actively applied. This characteristic is a crucial element of non-volatile memory, where information or states persist even in the absence of power or input stimuli.

To understand the behaviour of the circuit across different frequency ranges, sinusoidal signals with varying frequencies are applied. The corresponding hysteresis loops obtained are presented in Figs. 7(a) and 7(b). Fig. 7(a) displays the PHL curves obtained for frequencies ranging from 1kHz to 7kHz. In this figure, the curves are plotted with frequency varying in steps of 2kHz. Fig. 7(b) demonstrates the curves for frequency varying from 1MHz to 7MHz in steps of 2MHz. The capacitance of C_2 has been suitably modified to produce PHL curves with zero-crossing for the given range of frequencies. An additional crucial aspect of a meminductive device is the PHL lobes' area decreasing with frequency. This characteristic can be effectively visualized in the PHL curves shown in Fig. 7.

Additionally, to explore the maximum operating frequency of the proposed FOMI, the circuit underwent testing with high-frequency sinusoidal signals. Fig. 8 displays the results analyzed for an input signal of 10MHz frequency for different values of the fractional order parameter (α = 0.2, 0.5, and 0.8). As evident from the figure, PHL curves for all tested α values exhibit the dumbbell hysteresis shape, confirming the FOMI's functional operation even at a high frequency of 10 MHz. However, these curves are observed with slight shift in pinched point.

5.2 Simulation results of proposed floating FOMI in decremental configuration

The proposed floating FOMI presented in Fig. 4 is simulated in decremental mode. The circuit behaved satisfactorily, as shown by the results of the non-volatility test and transient analysis. The PHL curves observed for a sinusoidal signal with frequency varying from 100kHz to 7MHz are depicted in Fig. 9. All these curves have been plotted for α = 0.5. From Fig. 9, it can be visualized that the size of the lobes reduces as frequency increases, this

confirms the satisfactory behaviour of the proposed emulator in the complete range of frequencies for the decremental mode of configuration.

To analyze the effect of the variations in α on PHL curves, the circuit has been simulated for α varying from 0.2 to 0.8 under the stimulation of a sinusoidal voltage signal. The curves obtained are displayed in Fig. 10. The slight modifications in the shape of loop demonstrate the control of the fractional order in meminductance of the designed emulator.

5.3 Simulation results of the proposed grounded FOMI

In the presented grounded FOMI circuit illustrated in Fig. 3, a sinusoidal signal with a phase difference of 90^0 has been applied at the input. This sinusoidal input serves as a dynamic stimulus to assess the performance of the grounded FOMI across a range of frequencies. The frequency of the applied signal has been systematically varied, spanning from 100kHz to 7MHz. For the evaluation of the meminductive behaviour under these varying frequencies, the R and C values are configured according to the Foster-I form with a specific fractional order, in this case, α = 0.5. This choice of α is significant, as it represents a balanced fractional order that influences the circuit's meminductive characteristics.

The resulting responses of the grounded FOMI are captured in curves presented in both Fig. 11 and Fig. 12. Fig. 11 illustrates the behaviour in the incremental mode, where the meminductive response is observed during the increasing phase of the applied signal. On the other hand, Fig. 12 displays the response in the decremental mode, focusing on the meminductor's behaviour during the decreasing phase of the signal. These curves validate the proper functioning of the proposed grounded meminductor across the entire frequency range for both incremental and decremental configurations. The fact that the curves exhibit consistent and reliable behaviour throughout this frequency spectrum affirms the robustness and effectiveness of the grounded FOMI circuit.

5.4 Analysis of the proposed meminductor subjected to environmental conditions

In reality, a device will experience a variety of non-ideal environmental and manufacturing conditions, which will affect its behaviour. It is necessary to analyze a device's characteristics, when subjected to practical conditions, to determine its tolerance in a real-world environment. In this section, the impact of the variations of various parameters on the behaviour of the proposed FOMI has been investigated.

5.4.1 Temperature variation analysis

The performance characteristics of the proposed meminductor have been thoroughly investigated by examining its behaviour across a temperature range spanning from -50°C to +50°C. The analysis involves the generation of PHL curves for the floating FOMI under both incremental and decremental configurations. The outcomes of this investigation are represented in Fig. 13(a) for incremental mode and Fig. 13(b) for decremental mode.

In these figures, the meminductor's behaviour is assessed for $\alpha = 0.5$, a choice that represents a balanced configuration for the meminductor. The PHL curves plotted in these figures depict the slight deviations in the meminductive response under varying temperatures.

The observed slight deviations in the PHL curves do not compromise the overall operation of the proposed meminductor. Instead, they confirm the meminductor's ability to maintain functionality and stability across a defined temperature span. This resilience is indicative of the suggested meminductor's capability to withstand temperature variations without substantial impact on its performance, making it a reliable and robust component for applications that may be subjected to temperature fluctuations within the specified -50°C to +50°C range.

5.4.2 Supply voltage variations

The impact of supply voltage variations on the performance of the proposed FOMI has been systematically examined by subjecting the circuit to an approximate 10% fluctuation in the supply voltage. The resulting PHL curves, obtained by simulating the suggested meminductor with a sinusoidal signal of 100 kHz frequency and $\alpha = 0.5$, are presented in Fig. 14. The depicted waveforms in Fig. 14 illustrate the PHL curves in the φ vs. i plane, showcasing the meminductive response to variations in supply voltage. Specifically, the curves are generated for supply voltage fluctuations ranging from 0.81V to 0.99V for V_{DD} and from -0.81V to -0.99V for V_{SS}.

Despite the variations in supply voltage within the specified range, the suggested meminductor demonstrates satisfactory performance. The observed PHL curves reveal only slight modifications in lobe shape, indicating that the meminductive behavior remains stable even under supply voltage variations of 10%.

5.4.3 Variations due to capacitance tolerance:

During the fabrication process, achieving exact component values is often an impossible task. In reality, component values are inherently prone to certain tolerances or variations, which can potentially lead to circuit failure or suboptimal performance. Recognizing this inherent variability, an examination of the suggested meminductor's behaviour in the φ vs. i plane has been conducted by varying capacitor values. This investigation aims to understand how fluctuations in capacitance values influence the meminductive response.

Fig. 15 visually presents the PHL curves for the floating FOMI for both incremental and decremental configurations. These curves are generated for different values of capacitance, where C_2 is set at 5pF, 10pF, and 20pF. This variation in capacitance values provides insight into how the meminductive behaviour adapts to different component tolerances.

5.4.4 Monte-Carlo Analysis:

In the practical development of a MOSFET-based circuit, factors such as the aspect ratio and threshold voltage of the MOSFET can undergo variations due to fabrication imperfections. To assess the impact of these variations on the properties of the proposed meminductor, Monte Carlo (MC) analysis has been employed. The Gaussian distribution function has been used to perform the MC analysis with 5% variations in the MOSFET's aspect ratio and threshold voltage. Figs. 16 (a) and 16 (b) show the MC plots of PHL curves for a 100 kHz sinusoidal signal, considering threshold modifications for incremental and decremental meminductors, respectively. Each curve in these figures represents the outcome of the MC analysis for a specific run, with a total of 100 runs being conducted. These MC plots serve to demonstrate the robustness of the proposed circuit's meminductive behaviour across a range of mismatch scenarios.

The observed PHL curves in the MC plots affirm the circuit's ability to maintain meminductive functionality over the entire range of mismatch induced by variations in MOSFET specifications. Despite the inherent variations in the aspect ratio and threshold voltage, the form of the PHL curves remains consistent, with only minor modifications. This resilience is crucial for real-world applications where fabrication-induced variations are inevitable, ensuring that the proposed meminductor reliably preserves its intended behaviour despite variations in MOSFET properties.

5.5 Quantification of error in fractional-order capacitance using Foster-I form

As mentioned in Section 1, commercially available fractional capacitors are not readily accessible. Hence various continuous fraction expansions utilizing RC networks have been proposed in the literature to approximate their values. The s-domain equation that characterizes the impedance of a CPE is expressed as follows [15]:

$$Z(s) = \frac{1}{C_\alpha s^\alpha} \tag{27}$$

here, C_α represents normalized capacitance expressed in Farad/sec$^{(1-\alpha)}$ (F/sec$^{(1-\alpha)}$). The value of this capacitor which depends both on frequency and order, is represented as:

$$C_\alpha = C\omega_0^{1-\alpha} \tag{28}$$

where, C is the capacitance value for which the fractional capacitor is designed. The phase (θ) of the impedance $Z(s)$ of CPE is expressed as:

$$\theta = -\alpha\pi/2 \tag{29}$$

In this paper, a fifth-order Foster-I network has been employed to realize the fractional capacitor (C_α) required in the proposed design. There exists a trade-off between the network order and the exact value of fractional impedance. The 5th order network was chosen in this work, as it enables the attainment of a highly accurate value of fractional capacitor impedance across several decades centered around the central frequency.

In this work, a capacitance C=10pF centred at f_0 = 100 kHz ($|Z|$ = 159.15kΩ) has been chosen for the realization of the fractional capacitor through 5th order Foster-I form. The impedance of the fractional capacitor, implemented using the network illustrated in Fig. 3(b) with the component values specified in Table 3, have been examined. The resulting impedances for various values of alpha are plotted in Fig. 17(a). In this figure, the impedance corresponding for α = 0.2, 0.4, 0.6 and 0.8 have been plotted for a frequency range of 10^3 Hz to 10^7 Hz (4 decades). The waveforms depicting % error as compared to the impedance obtained through Equation (27) have been plotted in Fig. 17(b). These waveforms show that near centre frequency, the 5th order Foster-I form accurately realizes the fractional capacitance. The error in impedance magnitude is almost negligible in the frequency range of 20kHz to 500kHz and increases beyond this range.

The waveforms showing phase response of fractional capacitance realized through Foster-I RC network and the corresponding error related to different values of α have been plotted in Fig. 18(a) and 18(b) respectively. These plots have also been drawn for a frequency range of 4 decades (10^3 Hz to 10^7 Hz). Fig. 18(b) depicts that error in phase response is almost negligible for 2 decades of frequency variation (from 10kHz to 1MHz).

Figs. 17 and 18 confirm the validity of the 5^{th} order Foster-I RC network for implementing the desired fractional capacitor.

6 Applications of the suggested FOMI as a chaotic oscillator

This section introduces a chaotic oscillator implemented using the proposed FOMI. The incorporation of FOMI in chaotic oscillator design introduces additional degrees of freedom, resulting in more intricate and diverse dynamical behavior. The fractional order, denoted by 'α', offers precise control over the system's dynamics, facilitating customization of the oscillator's behavior to meet specific requirements [14, 16]. Despite the advantages of fractional-order chaotic oscillators, including richer dynamics and enhanced controllability, their implementation entails increased complexity and challenges.

In the suggested implementation of Chua's oscillator, the fractional order model is exclusively applied to the meminductor flux characteristic, while the remaining states are maintained as integer order. Fig. 19 illustrates the basic configuration of Chua's oscillator, where the traditional nonlinear element (M_L) is replaced with the proposed VDIBA and CF based meminductor. The oscillator includes one inductor (L), two capacitors (C_1 & C_2), a resistor (R_1), an OPAMP-based negative impedance converter (-R_2), and one meminductor (M_L).

The four state variables of this oscillator are: current through the inductor (I_L), the capacitor node voltages (V_X & V_Y), and flux (φ_{MI}) and integral of flux (ρ_{MI}) of meminductor. This meminductor-based oscillator's state space dynamics can be expressed with the help of 1st order differential equations as shown below:

$$\left.\begin{aligned} C_A\frac{dV_X}{dt} &= \frac{V_X}{R_B} - I_L \\ C_B\frac{dV_Y}{dt} &= I_L - I_{ML} \\ L\frac{dI_L}{dt} &= V_X - V_Y - I_L\cdot R_B \\ \frac{d\varphi_{MI}}{dt} &= V_Y \\ \frac{d\rho_{MI}}{dt} &= \varphi_{MI} \end{aligned}\right\} \qquad (30)$$

In above analysis, C_A, C_B, L, R_A and R_B represent the values of the respective elements (capacitance, inductance, or resistance) in the chaotic oscillator. To model the current through the meminductor, the equation $I_{LM} = (A+B\rho_{MI})\varphi_{MI}$, as derived in Equation (21), can be employed. In this equation, α is considered to be equal to unity for the sake of simplicity (α = 1 for an ideal capacitor). Under this assumption, the parameters A and B are characterized by Equation (31).

$$\left.\begin{aligned} A &= \frac{\gamma \cdot \beta}{R_1 \cdot C_\alpha} \frac{K}{2\sqrt{2}} \left(V_{SS} + 2V_{th}\right) \\ B &= \frac{\gamma^2 \cdot \beta^2}{R_1^2 \cdot C_2 \cdot C_\alpha^2} \cdot \frac{K \cdot G_m}{2\sqrt{2}} \end{aligned}\right\} \tag{31}$$

A set of equilibrium points can be derived by setting all derivatives of Equation (30) equal to zero. This set of equilibrium points considering, $\dot{V}_X, \dot{V}_Y, \dot{I}_L, \dot{\varphi}_{MI}, \dot{\rho}_{MI} = 0$, is defined as:

$$E = \left\{ \left(V_X, V_Y, I_L, \varphi_{MI}, \rho_{MI}\right) | V_X = V_Y = I_L = \varphi_{MI} = 0, \rho_{MI} = m \right\} \tag{32}$$

where, *m* is a constant real number. At equilibrium point given by Equation (32), the Jacobian matrix (*J*) is defined as:

$$J = \begin{bmatrix} \frac{1}{C_A R_B} & 0 & -\frac{1}{C_A} & 0 & 0 \\ 0 & 0 & \frac{1}{C_B} & -\frac{1}{C_B}\left(A + Bm\right) & 0 \\ \frac{1}{L} & -1 & \frac{-R_A}{L} & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{33}$$

The dynamics of the Chua's oscillator (Fig. 19) can be analyzed by finding the characteristic equation associated with the Jacobian matrix expressed Equation (33). This equation which gives the eigenvalues of the matrix is defined as:

$$Det\left(J - \lambda I\right) = 0 \tag{34}$$

here, λ represents the eigenvalues and are crucial for determining the stability of the Chua's oscillator at the equilibrium point specified by Equation (32). Using MATLAB, the eigenvalues of the system are calculated as:

$$\lambda^5 + a_1\lambda^4 + a_2\lambda^3 + a_3\lambda^2 + a_4\lambda = 0 \tag{35}$$

here, coefficients a_1, a_2, a_3, and a_4 are defined by Equation (36).

$$\left.\begin{aligned} a_1 &= \frac{R_A}{L} - \frac{1}{C_A R_B} \\ a_2 &= \frac{1}{C_B}\left(1 + A + Bm\right) + \frac{1}{C_A L}\left(1 - R_A \big/ R_B\right) \\ a_3 &= \frac{1}{C_B}\left\{\left(A + Bm\right)\left(\frac{R_A}{L} - \frac{1}{C_A R_B}\right) - \frac{1}{C_A R_B}\right\} \\ a_4 &= \frac{1}{C_A C_B L}\left(A + Bm\right)\left(1 - R_A \big/ R_B\right) \end{aligned}\right\} \tag{36}$$

Using the characteristic equation given by Equation (34), the eigen values can be obtained for the chaotic circuit of Fig. 19. From Equation (35), it can be easily realized that Jacobian matrix '*J*' has a zero value and four nonzero values. Since, the coefficients a_1, a_2, a_3, and a_4 are all nonzero, if the real parts of all eigenvalues of Equation (35) (except $\lambda = 0$) are negative, the system is stable. As per Routh–Hurwitz criterion the system is stable only when, $\Delta_1 = a_1$, $\Delta_2 = (a_1a_2 - a_3)/a_1$, $\Delta_3 = a_3-(a_1^2a_4)/(a_1a_2 - a_3)$, and $\Delta_4 = a_4$, all are positive. Otherwise, the system is unstable, and chaotic response can be generated. Fractional-order systems exhibit significantly more complex behavior than their integer-order counterparts due to the challenging nature of fractional-order calculus. The previous analysis was conducted assuming an ideal capacitor with $\alpha = 1$. Interestingly, when considering a fractional capacitor ($0 < \alpha < 1$), the chaotic behavior remains at least as stable as the integer-order system [17]. However, maintaining chaos in the fractional system depicted in Fig. 19 requires the same number of eigenvalues to remain within the unstable region. The essential condition to achieve unstable eigenvalues is outlined below:

$$\alpha > \frac{2}{\pi}\tan^{-1}\left(\frac{\left|\mathrm{Im}(\lambda)\right|}{\mathrm{Re}(\lambda)}\right) \tag{37}$$

The behavior of the chaotic oscillator, as illustrated in Fig. 19 and implemented with the recommended fractional-order meminductor with $\alpha = 0.5$, has been explored through LTspice simulations. The negative resistance ($-R_2$) is achieved by the OPAMP-based negative impedance converter when $R_a = R_b = 2$k. The rest of the elements are set to have the following values: L=120mH, C_1=65nF, C_2=10nF, and R_1=200. Fig. 20 shows the 2-D projection plots that are observed for the different state variables. These plots validate the chaotic dynamics exhibited by the system when utilizing fractional values of α.

7 Conclusion

In this paper, the feasibility of designing grounded/floating FOMI based on a few active blocks has been investigated. VDIBA and CF blocks have been selected for their straightforward construction, differential capabilities, and inherent tuning options. Verification tests, including meminductor fingerprints, PHL with zero-crossing, and non-volatility tests, have all been employed to ensure the intended functionality of the proposed circuit. The circuit's ability to operate in both incremental and decremental modes was demonstrated using a simple switch connected to the input terminals. LTspice simulation results indicate the circuit's effectiveness up to 7MHz. Temperature variations (-50°C to

+50°C), capacitance changes (5pF to 20pF), and supply voltage fluctuations (10%) were considered in PHL curves, affirming the circuit's practicality in real-world scenarios. Additionally, the proposed fractional-order meminductor's realization was validated through a Monte Carlo study. The paper illustrates the successful implementation of a chaotic oscillator using the suggested fractional-order floating meminductor emulator.

References

1. Tsirimokou, G., Psychalinos, C., and Elwakil, A. *Design of CMOS Analog Integrated Fractional-Order Circuits* (2017). DOI: https://doi.org/10.1007/978-3-319-55633-8.
2. Dutta Roy, S. C. "On the Realization of a Constant-Argument Immittance or Fractional Operator", *IEEE Trans. Circuit Theory*, **14**(3), pp. 264–274 (1967). DOI: 10.1109/TCT.1967.1082706
3. Aroudi, A. El, Rodriguez, E., and Orabi, M. "Modeling of switching frequency instabilities in buck- based DC – AC H-bridge inverters", *Int. J. Circuit Theory Appl.*, **39**(2), pp. 175-193 (2010). DOI: 10.1109/MEPCON.2008.4562377
4. Chua, L. "The missing memristor", *IEEE Trans. Circuit Theory*, **C**(5), pp. 507–519 (1971). DOI: 10.1109/TCT.1971.1083337
5. Strukov, D. B., Snider, G. S., Stewart, D. R., et al. "The missing memristor found", *Nature*, **453**(7191), pp. 80–83 (2008). DOI: https://doi.org/10.1038/nature06932
6. Khalil, N. A., Said, L. A., Radwan, A. G., et al. "Emulation circuits of fractional-order memelements with multiple pinched points and their applications", *Chaos, Solitons and Fractals*, **138**, pp. 109882 (2020). DOI: https://doi.org/10.1016/j.chaos.2020.109882
7. Khalil, N. A., Said, L. A., Radwan, A. G., et al. "General fractional order mem-elements mutators", *Microelectronics J.*, **90**, pp. 211–221 (2019). DOI: https://doi.org/10.1016/j.mejo.2019.05.018
8. Khalil, N. A., Said, L. A., Radwan, A. G., et al. "A universal floating fractional-order elements/memelements emulator", *NILES 2019 - Nov. Intell. Lead. Emerg. Sci. Conf.*, **1**, pp. 80–83 (2019). DOI: 10.1109/NILES.2019.8909296
9. Khalil, N. A., Fouda, M. E., Said, L. A., et al. "A Universal Fractional-Order Memelement Emulation Circuit", *NILES 2019 - Nov. Intell. Lead. Emerg. Sci. Conf.*, **1**, pp. 67–70 (2019). DOI: 10.1109/NILES.2019.8909307
10. Gupta, A., Rai, S. K., and Gupta, M. "A Fractional-Order Meminductor Emulator Using OTA and CDBA with Application in Adaptive Learning Circuit", *Wirel. Pers.*

Commun., **131**(4), pp. 2675–2696 (2023). DOI: https://doi.org/10.1007/s11277-023-10566-2

11. Li, Y., Xie, L., Zheng, C., et al. "Modeling and hardware implementation of universal interface-based floating fractional-order mem-elements", *Chaos*, **33**(1) (2023). DOI: 10.1063/5.0124793
12. Petráš, I. and Chen, Y. Q. "Fractional-order circuit elements with memory", *Proc. 2012 13th Int. Carpathian Control Conf. ICCC 2012*, pp. 552–558 (2012). DOI: 10.1109/CarpathianCC.2012.6228706
13. Oresanya, B. O., Si, G., Xu, X., et al. "Conditions for realization of multiple pinch-off points in generalized fractional-order memory elements", *Int. J. Circuit Theory Appl.*, **51**(11), pp. 5050–5065 (2023). DOI: https://doi.org/10.1002/cta.3690
14. Khalil, N. A., Said, L. A., Radwan, A. G., et al. "General fractional order mem-elements mutators", *Microelectronics Journal*, **90**, pp. 211-221 (2019). DOI: https://doi.org/10.1016/j.mejo.2019.05.018
15. Tsirimokou, G. "A systematic procedure for deriving RC networks of fractional-order elements emulators using MATLAB", *AEU - Int. J. Electron. Commun.*, **78**, pp. 7–14 (2017). DOI: https://doi.org/10.1016/j.aeue.2017.05.003
16. Akgul, A. "Chaotic Oscillator Based on Fractional Order Memcapacitor", *J. Circuits, Syst. Comput.*, **28**(14), pp. 1950239 (2019). DOI: https://doi.org/10.1142/S0218126619502396
17. Tavazoei, M. S., and Haeri, M. "A necessary condition for double scroll attractor existence in fractional-order systems", *Physics Letters A*, **367**(1-2), pp. 102-113 (2007). DOI: https://doi.org/10.1016/j.physleta.2007.05.081

List of Figure Captions:

Fig. 1 VDIBA (a) Symbol (b) Circuit diagram

Fig. 2 CF (a) Symbol (b) Circuit diagram

Fig. 3 Proposed Grounded FOMI (a) Circuit diagram (b) Foster-I realization of fractional capacitor for $\alpha = 0.5$

Fig. 4: Proposed Floating Decremental/Incremental FOMI

Fig. 5 Transient analysis for the floating FOMI incremental configuration for a sinusoidal frequency of 100kHz for $\alpha = 0.5$

Fig. 6 Non-volatile behavior of the proposed floating FOMI for $\alpha = 0.5$

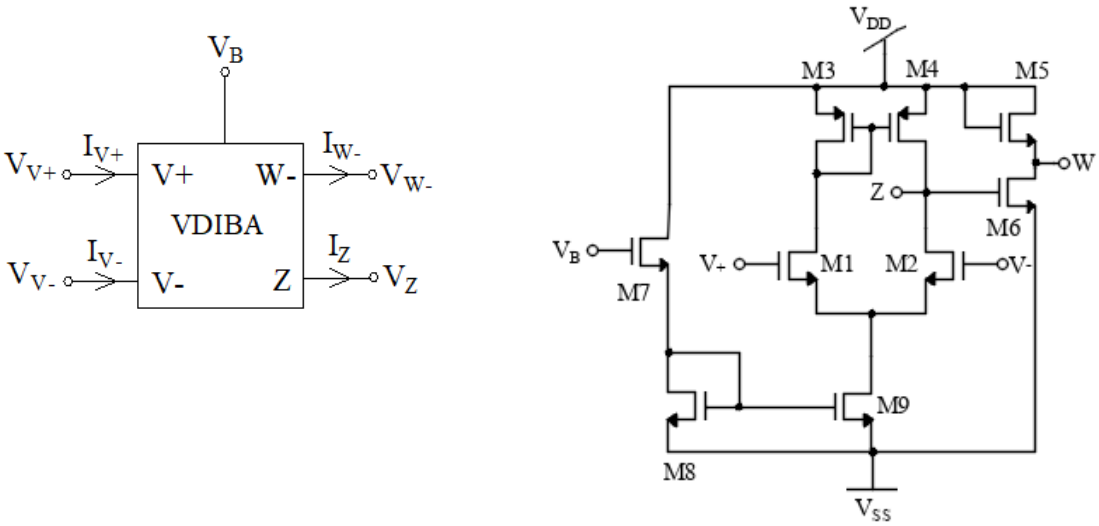
Fig. 7 PHL curves observed for the floating incremental FOMI for $\alpha = 0.5$ for a frequency range of (a) 1kHz to 7kHz (b) 1MHz to7MHz

Fig. 8 PHL curves observed for 10MHz sinusoidal signal for α = 0.2, 0.5 and 0.8
Fig. 9 PHL curves observed for the proposed floating decremental FOMI with α = 0.5 for a frequency range of (a) 100kHz to 700kHz (b) 1MHz to 7MHz
Fig. 10 PHL curves observed for decremental floating FOMI for α = 0.2 to 0.8
Fig. 11 PHL curves observed for the incremental FOMI with α = 0.5 for a frequency range of (a) 100kHz to 700kHz (b) 1MHz to7MHz
Fig. 12 PHL curves observed for the grounded decremental FOMI with α = 0.5 for a frequency range of (a) 100kHz to 700kHz (b) 1MHz to7MHz
Fig. 13 PHL curves of proposed floating FOMI (a) incremental (b) decremental configuration for temperature fluctuations from -50^0C to 50^0C
Fig. 14 PHL curves of the proposed FOMI for (a) floating incremental (b) floating decremental observed with ±10% deviations in supply voltage
Fig. 15. PHL curves of the proposed floating FOMI observed for deviations in C_2 for (a) incremental (b) decremental configuration
Fig. 16 PHL curves observed with MC analysis for the proposed floating FOMI for 5% variations in aspect ratio and threshold voltage for (a) incremental (b) decremental.
Fig. 17 Waveform corresponding to α = 0.2, 0.4, 0.6 and 0.8 for (a) Magnitude of impedance realized using 5th order Foster-I form (b) % error in the magnitude of fractional capacitor impedance
Fig. 18 Waveform corresponding to α = 0.2, 0.4, 0.6 and 0.8 for (a) Phase response of impedance realized using 5^{th} order Foster-I form (b) % error in the phase response of fractional capacitor impedance
Fig. 19. A fourth-order Chua's chaotic oscillator
Fig. 20 Projection plots of chaotic oscillator observed between space variables (a) V_X & I_L (b) V_X & I_{ML} (c) V_Y & V_X (d) V_Y & I_L

List of Table Captions:

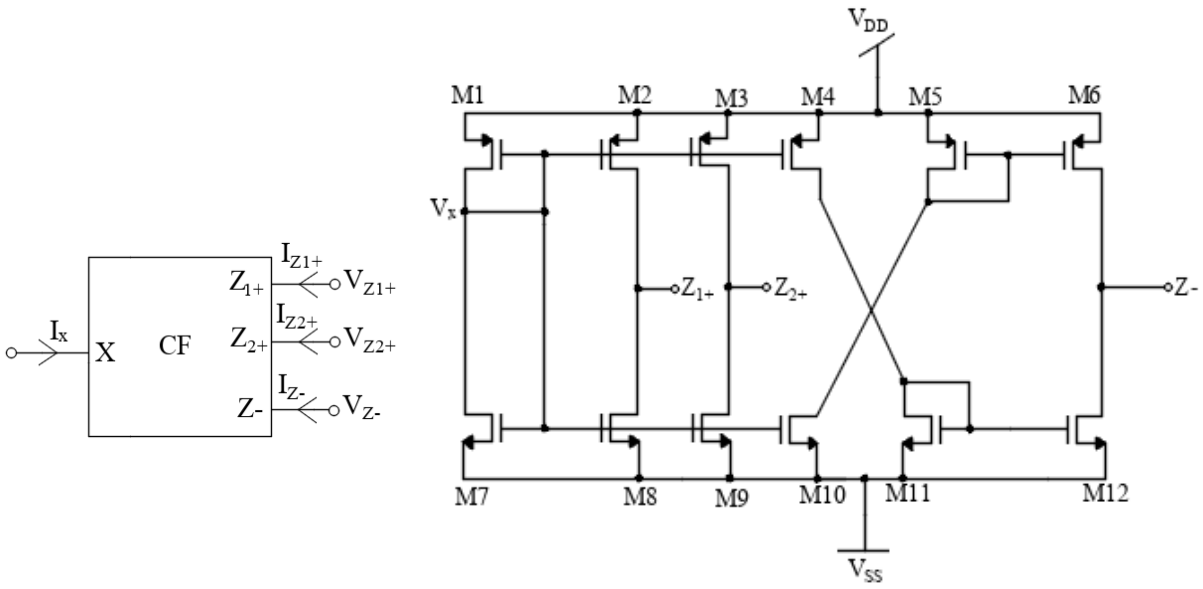
Table 1 Comparison of the suggested FOMI with existing meminductor emulators
Table 2: Aspect Ratios of MOSFETs
Table 3: R and C values for Foster-I form for various values of α

Figures:



(a) (b)

Figure 2. VDIBA (a) Symbol (b) Circuit diagram



(a) (b)

Figure 2. CF (a) Symbol (b) Circuit diagram

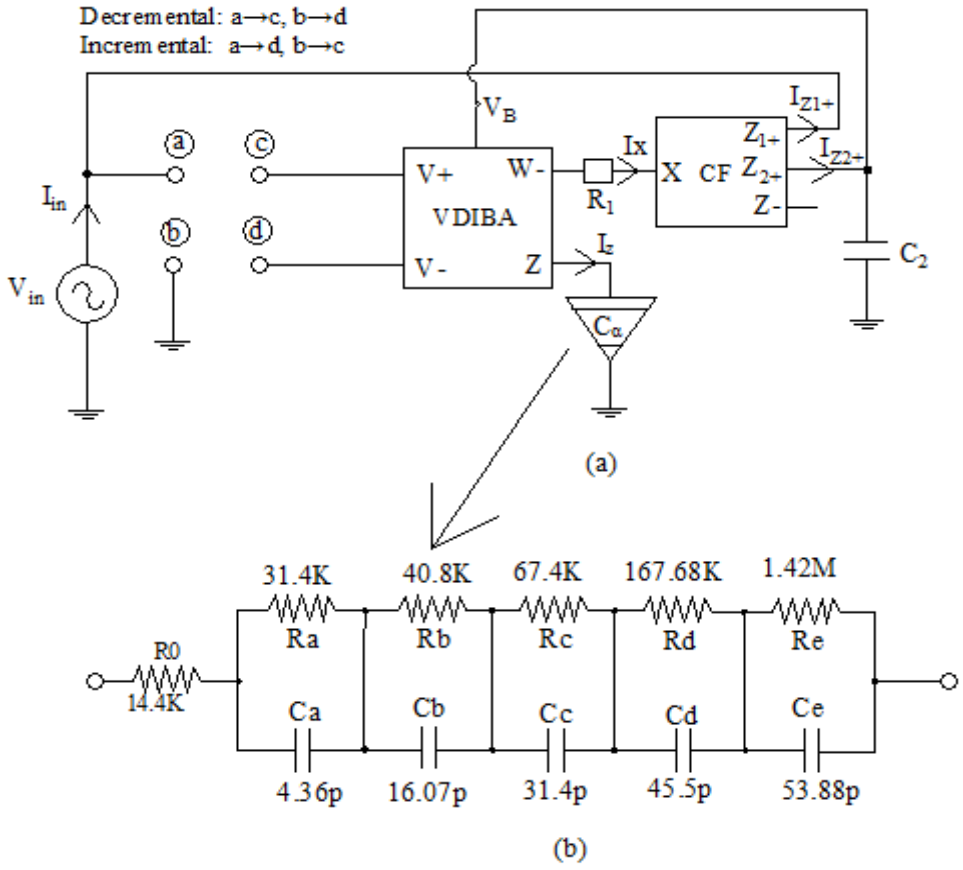


Figure 3. Proposed Grounded FOMI (a) Circuit diagram (b) Foster-I realization of fractional capacitor for $\alpha = 0.5$

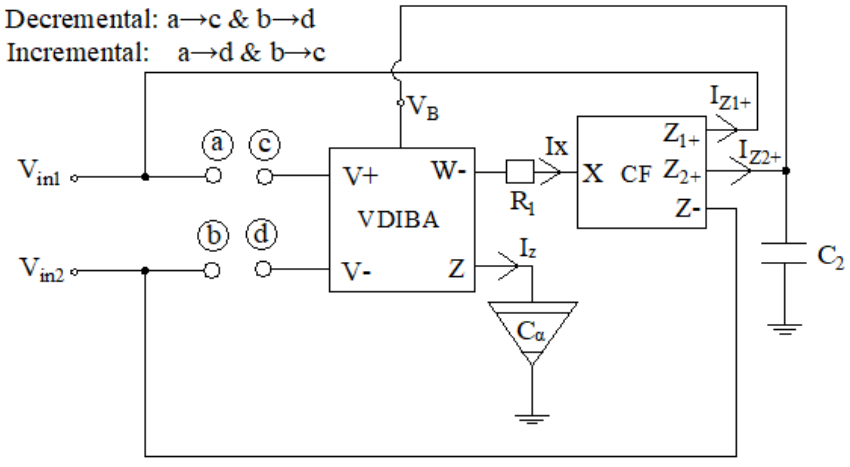


Figure 4. Proposed Floating Decremental/Incremental FOMI

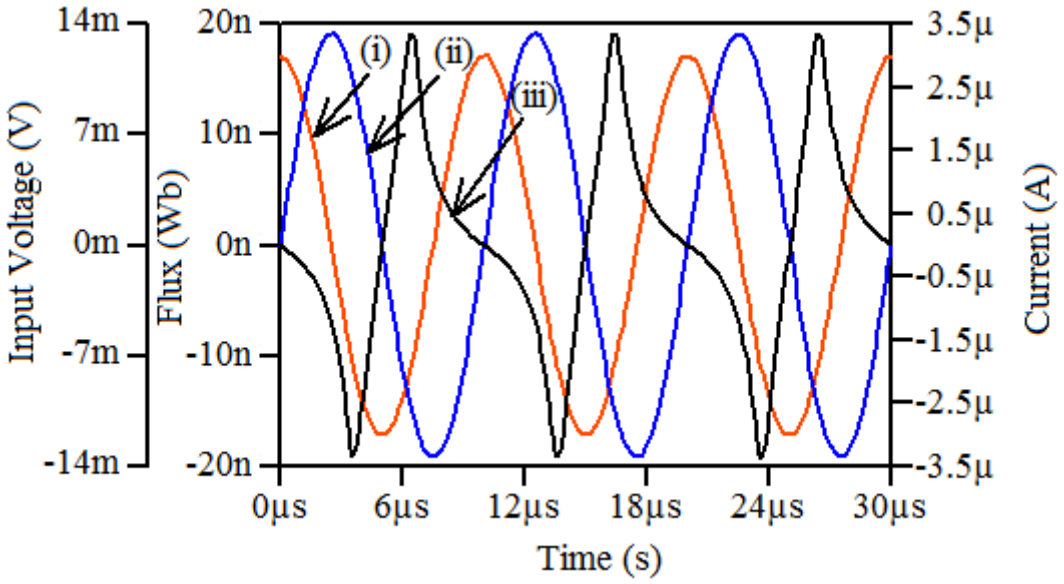


Figure 5. Transient analysis for the floating FOMI incremental configuration for a sinusoidal frequency of 100kHz for $\alpha = 0.5$

**for color reproduction on the Web*

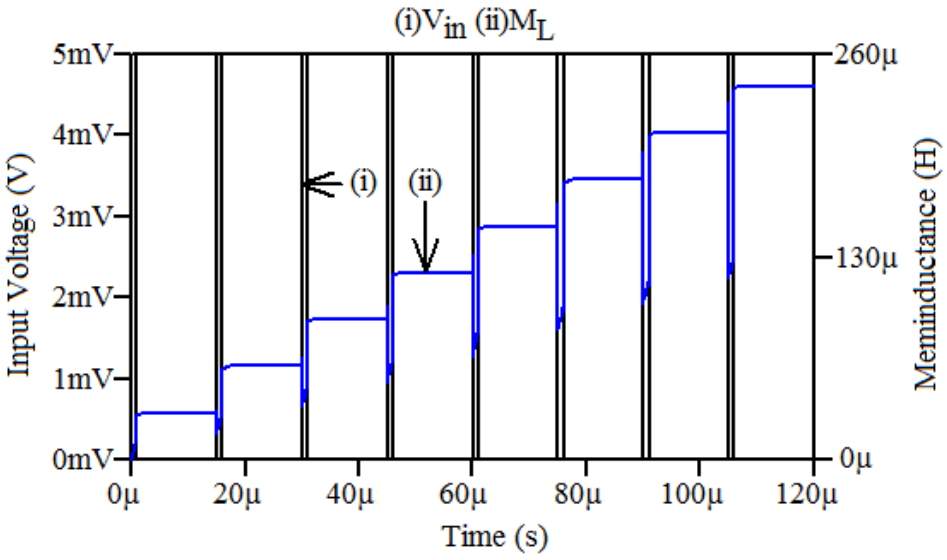
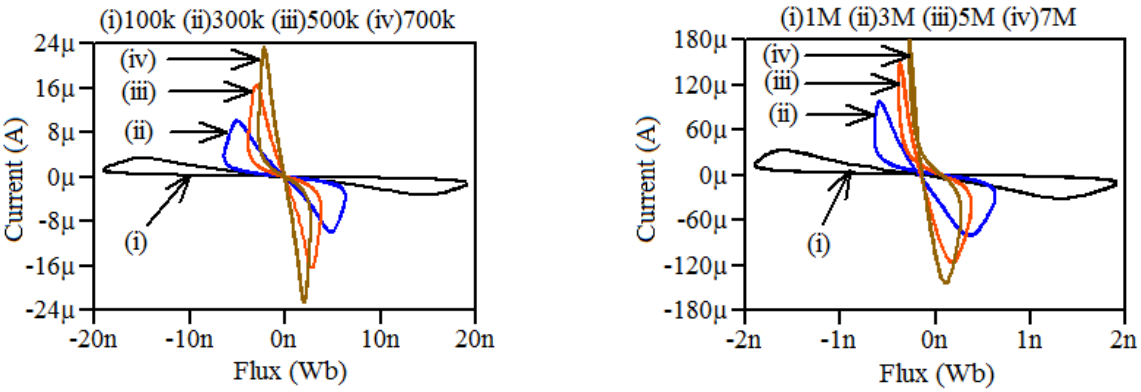


Figure 6. Non-volatile behavior of the proposed floating FOMI for $\alpha = 0.5$

**for color reproduction on the Web*



(a) (b)

Figure 7. PHL curves observed for the floating incremental FOMI for α = 0.5 for a frequency range of (a) 1kHz to 7kHz (b) 1MHz to7MHz

**for color reproduction on the Web*

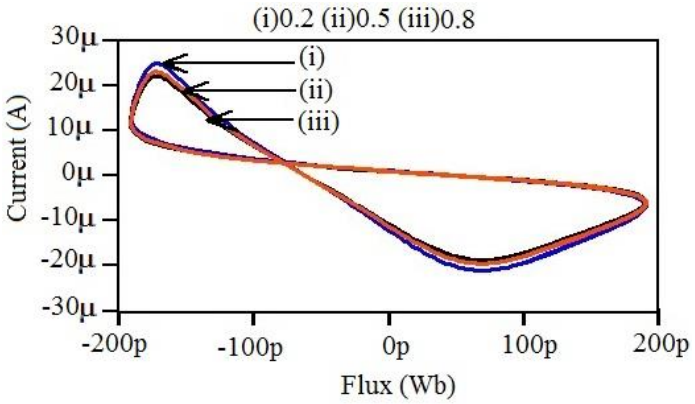
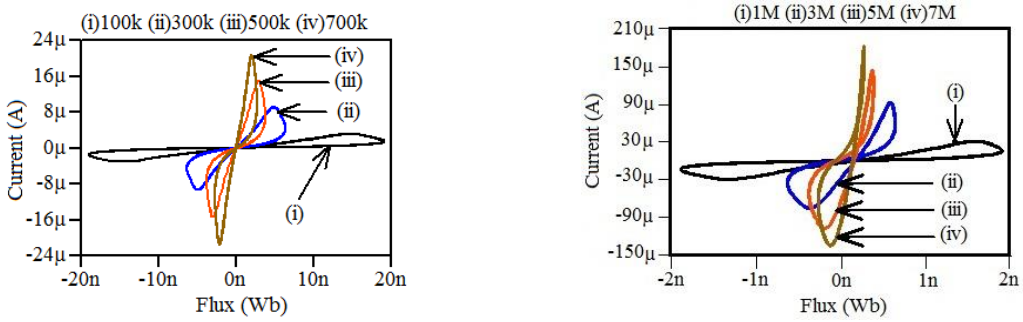


Figure 8. PHL curves observed for 10MHz sinusoidal signal for α = 0.2, 0.5 and 0.8

**for color reproduction on the Web*



(a) **(b)**

Figure 9. PHL curves observed for the proposed floating decremental FOMI with α = 0.5 for a frequency range of (a) 100kHz to 700kHz (b) 1MHz to 7MHz

**for color reproduction on the Web*

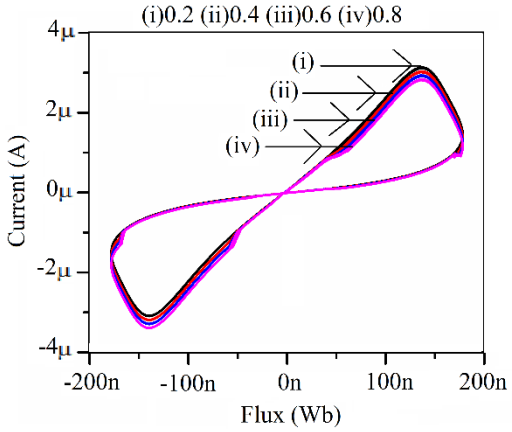
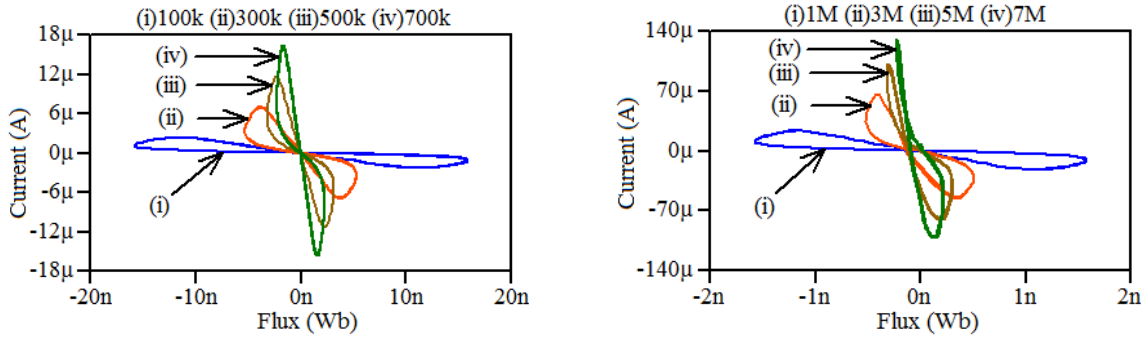


Figure 10. PHL curves observed for decremental floating FOMI for α = 0.2 to 0.8

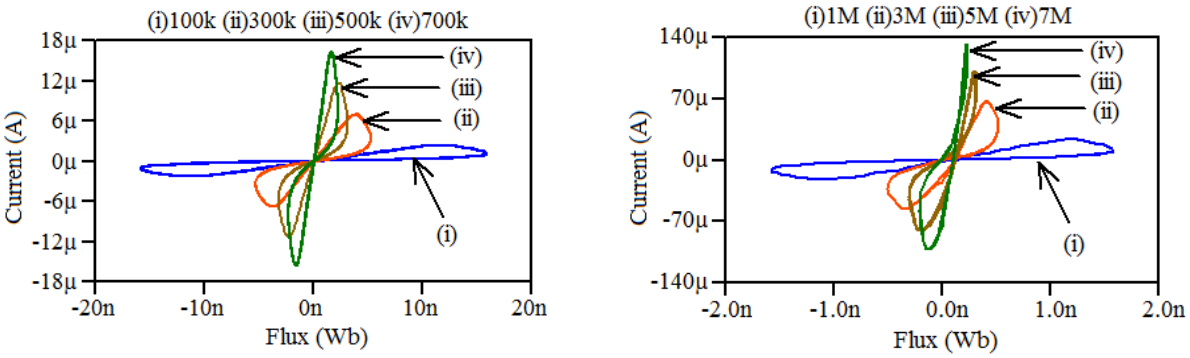
**for color reproduction on the Web*



(a) **(b)**

Figure 11. PHL curves observed for the incremental FOMI with α = 0.5 for a frequency range of (a) 100kHz to 700kHz (b) 1MHz to7MHz

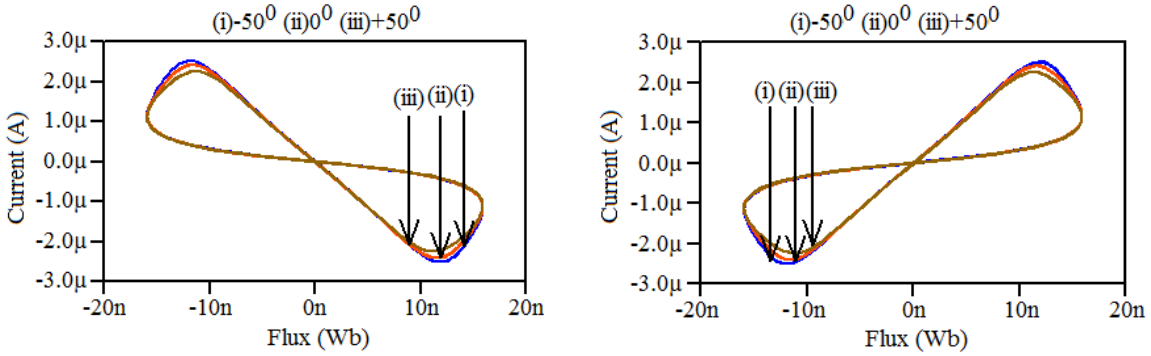
**for color reproduction on the Web*



(a) **(b)**

Figure 12. PHL curves observed for the grounded decremental FOMI with α = 0.5 for a frequency range of (a) 100kHz to 700kHz (b) 1MHz to7MHz

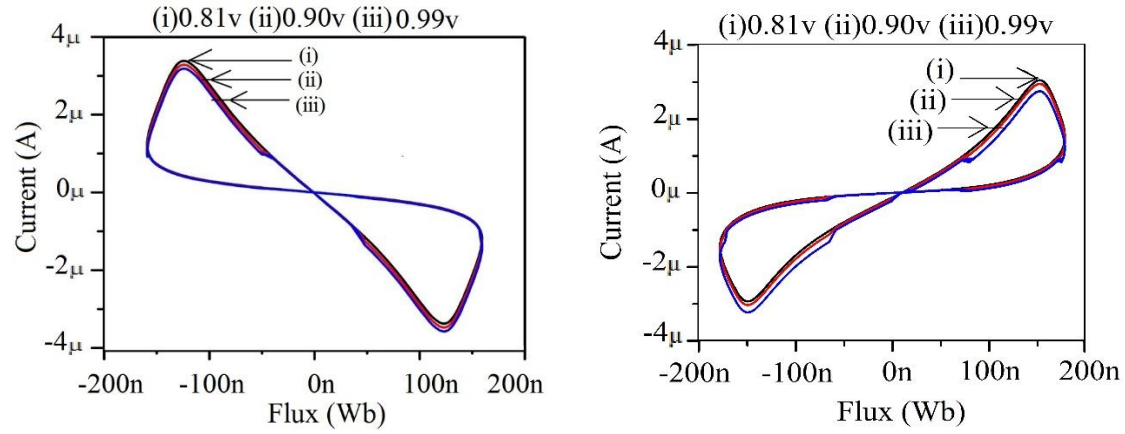
**for color reproduction on the Web*



(a) **(b)**

Figure 13. PHL curves of proposed floating FOMI (a) incremental (b) decremental configuration for temperature fluctuations from -50^{0}C to 50^{0}C

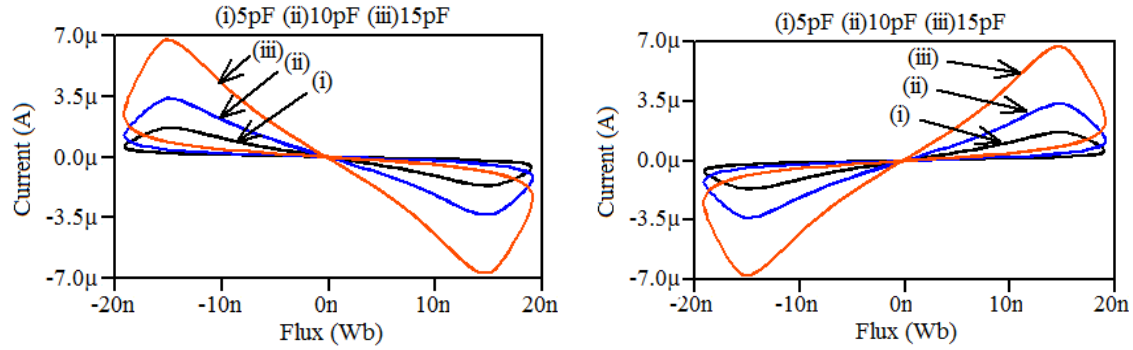
**for color reproduction on the Web*



(a) (b)

Figure 14. PHL curves of the proposed FOMI for (a) floating incremental (b) floating decremental observed with ±10% deviations in supply voltage

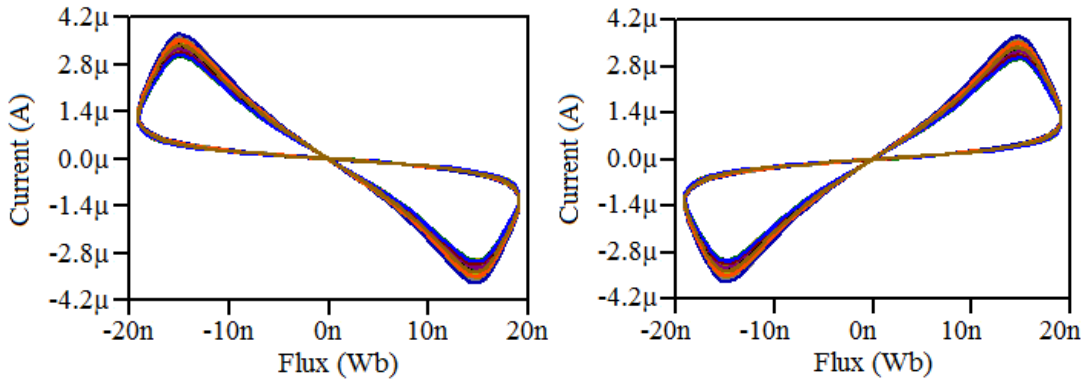
**for color reproduction on the Web*



(a) **(b)**

Figure 15. PHL curves of the proposed floating FOMI observed for deviations in C_2 for (a) incremental (b) decremental configuration

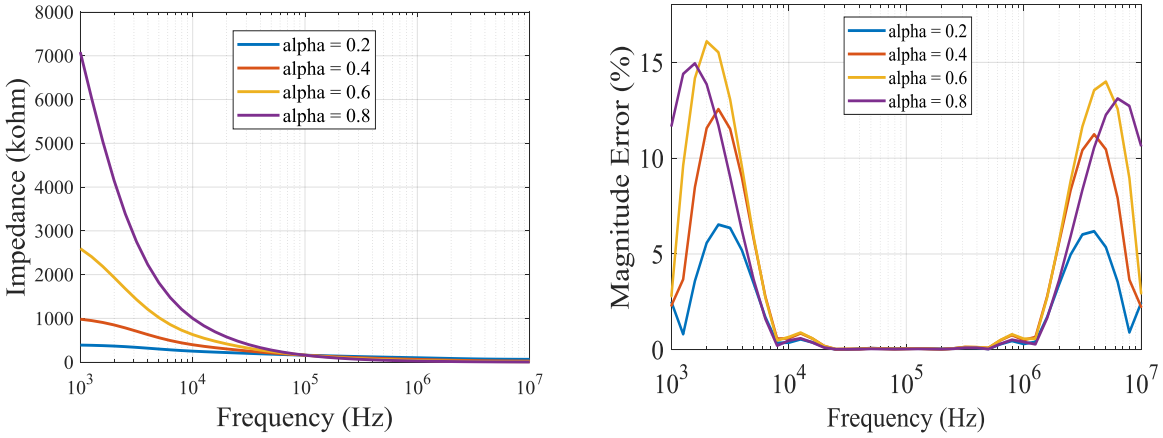
**for color reproduction on the Web*



(a) (b)

Figure 16. PHL curves observed with MC analysis for the proposed floating FOMI for 5% variations in aspect ratio and threshold voltage for (a) incremental (b) decremental.

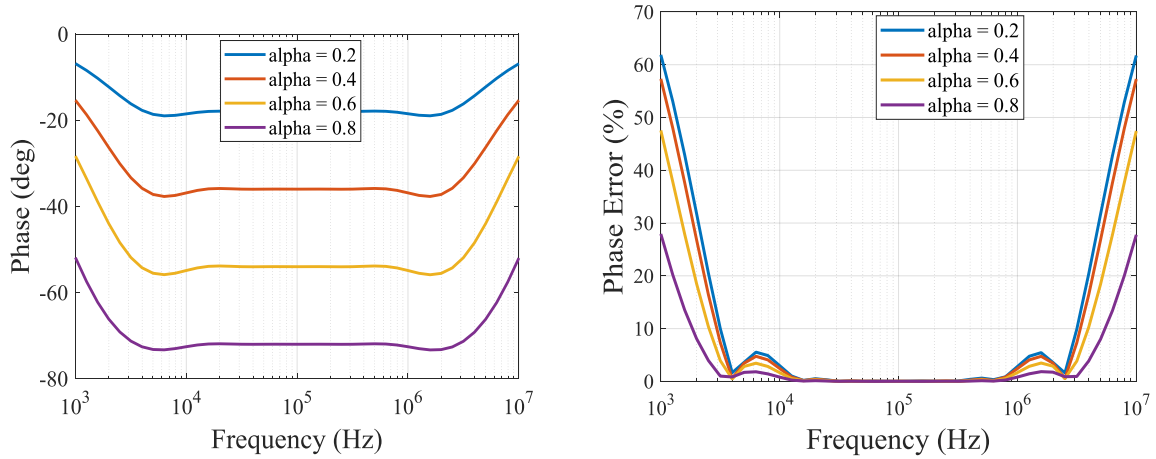
**for color reproduction on the Web*



(a) (b)

Figure 17. Waveform corresponding to α = 0.2, 0.4, 0.6 and 0.8 for (a) Magnitude of impedance realized using 5th order Foster-I form (b) % error in the magnitude of fractional capacitor impedance

**for color reproduction on the Web*



(a) (b)

Figure 18. Waveform corresponding to α = 0.2, 0.4, 0.6 and 0.8 for (a) Phase response of impedance realized using 5th order Foster-I form (b) % error in the phase response of fractional capacitor impedance

**for color reproduction on the Web*

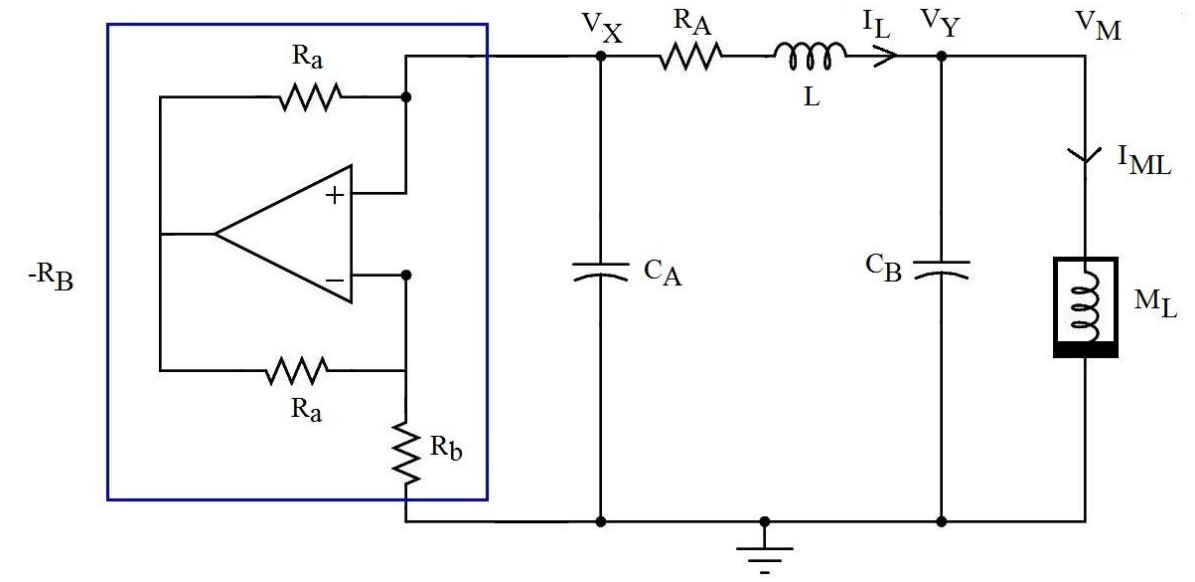


Figure 19. A fourth-order Chua's chaotic oscillator

**for color reproduction on the Web*

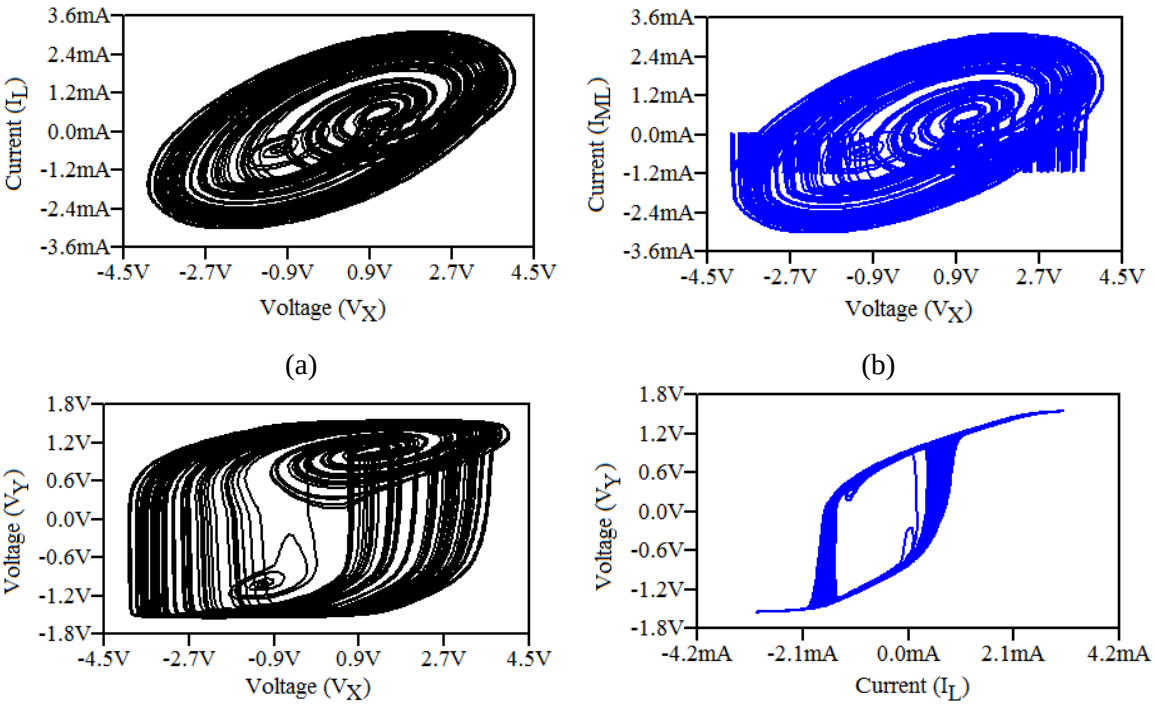


Figure 20. Projection plots of chaotic oscillator observed between space variables (a) V_X & I_L (b) V_X & I_{ML} (c) V_Y & V_X (d) V_Y & I_L

**for color reproduction on the Web*

Tables:

Table 1 Comparison of the suggested FOMI with existing meminductor emulators

Ref No	No. of ABB	No. of R	No.of C	No.of L	Multiplier Less	Memristorless	Operating Freq.	Electronic Tunability	*F/G
[6]	5-CCII	4	0	0	No	Yes	10kHz	No	F+G
[7]	2 CCII+ 1 TA, 1 BOCCII	0	0	0	Yes	No	2kHz	No	F
[8]	1 DVCC+ 1 BOCCII,	1	0	0	Yes	No	2kHz	No	F
[9]	3-CCII	3	1	0	Yes	Yes	10Hz	No	F
[10]	2 OTA+ 1 CDBA	0	1	0	Yes	Yes	3MHz	No	F
[11]	4-AD844+ 1-AD633	2	0	1	No	Yes	5kHz	No	F
[12]	4-OPAMP	5	1	1	Yes	No	---	No	G
[13]	3-CCII	2	0	0	No	Yes	500Hz	NO	G
Our work	**1-VDIBA+ 1-CF**	**0**	**1**	**0**	**Yes**	**Yes**	**7MHz**	**No**	**G+F**

CCII: Second generation current conveyor, TA: Transconductance amplifier, BOCCII: Balanced Output CCII, DVCC: Differential Voltage Current Conveyor, OTA: Operational transconductance amplifier, CDBA: Current differencing buffered amplifier, OPAMP: Operational amplifier

G: Grounded; F: Floating

Table 2: Aspect Ratios of MOSFETs

VDIBA			CF		
MOS Transistors	W (µm)	L (µm)	MOS Transistors	W (µm)	L (µm)
M1-M9	54	0.18	M10-M15	14.04	1.08
			M16-M21	56.16	1.08

Table 3: R and C values for Foster-I form for various values of α

Alpha (α)	$R_0(\Omega)$	$R_a(\Omega)$	$R_b(\Omega)$	$R_c(\Omega)$	$R_d(\Omega)$	$R_e(\Omega)$	C_a(F)	C_b(F)	C_c(F)	C_d(F)	C_e(F)
0.2	63.43k	40.89k	31.35k	34.73k	54.72k	174.23k	2.44p	17.37p	51.34p	111.55p	247.96p
0.3	3.95k	4.22k	3.89k	4.93k	8.98k	38.05k	2.65p	14.90p	38.27p	72.87p	134.32p
0.4	24.23k	38.1k	41.92k	60.61k	128.60k	751.68k	3.26p	14.74p	33p	54.9p	82.23p
0.5	14.4k	31.4k	40.8k	67.4k	167.68k	1.42M	4.36p	16.07p	31.4p	45.50p	53.88p
0.6	8.29k	23.92k	36.54k	68.83k	201.83k	2.71M	6.30p	19.11p	32.64p	40.99p	36.85p
0.7	4.45k	1.65k	2.95k	6.36k	22.11k	535.25k	9.94p	25.06p	37.45p	40.54p	25.95p
0.8	2.11k	10k	20.72k	50.98k	213.87k	11.66M	17.97p	37.97p	49.64p	46.03p	18.64p

Biographies

Shireesh Kumar Rai received the B. Tech. degree in electronics and communication engineering from the Noida Institute of Engineering and Technology, Gr. Noida in 2008, the M.Tech. degree in VLSI Design from the YMCA University of Science and Technology, Faridabad, in 2012, and the Ph.D. degree from the Netaji Subhas Institute of Technology (University of Delhi), New Delhi, in 2017. From 2008 to 2010, he was a Lecturer at the Galgotias College of Engineering and Technology, Gr. NOIDA. From 2012 to 2013, he was a Sr. Lecturer at Echelon Institute of Technology, Faridabad. In 2013, he joined as a Teaching Cum Research Fellow with the Netaji Subhas Institute of Technology (University of Delhi). Since 2017, he has been an Assistant Professor with the Thapar Institute of Engineering and Technology, Patiala. He has authored/co-authored 43 research papers in SCI journals, 5 research papers in Scopus journals, and 6 research papers in international conferences. His research areas include mem-elements emulator circuits, chaotic circuits for cryptographic applications, analog integrated circuits for signal processing, and bandgap reference circuits.is currently working as Associate Professor, ECE Department, Netaji Subhas University of

Bhawna Aggarwal Technology, Dwarka, New Delhi, India. She has a total work experience of 20 years and received her PhD. from NIT Kurukshetra, Haryana, India in the year 2016. She received her M.E. in Electronics & Communication Engineering from Delhi College of Engineering, New Delhi in 2006 and B.Tech. in Electronics & Communication Engineering from Indira Gandhi Institute of Technology, Delhi in 2002. Low power design techniques, Mem-element circuits and Fractional order filters are her areas of research. She is a Senior Member of IEEE and an active member of IETE, ISTE and Vibha Societies. She has published 30+ research papers in SCI/SCIE journals and more than 30 papers in reputed international conferences.

Rupam Das received his B.Tech. Degree in Electronics & Communication Engineering in 2006 and M.Tech. Degree in VLSI & Microelectronics in 2008 from West Bengal University of Technology, West Bengal, India. He has completed his Ph.D from Indian Institute of Technology (Indian School of Mines), Dhanbad, India in 2021. Currently he is working as an Assistant Professor in Electronics and Communication Engineering at Asansol Engineering College, Asansol, India. He has published many papers in the reputed international journals and conferences. His main research interest is in the area of current mode analog integrated circuit design and mixed signal circuit design.

Journal of Optics
https://doi.org/10.1007/s12596-024-01954-3

RESEARCH ARTICLE

A Graded Index Hybrid Photonic Crystal Fiber for Supercontinuum Generation using $AsSe_2$- As_2S_5

Drishti Singh Tomer[1] · **Ajeet Kumar**[1]

Received: 24 April 2024 / Accepted: 30 May 2024
© The Author(s), under exclusive licence to The Optical Society of India 2024

Abstract
We design and simulate an especially engineered photonic crystal structure. $AsSe_2$ as the solid core and As_2S_5 in graded-index hybrid cladding. Hybrid cladding region with inner cladding as a graded helical structure and outer cladding is a hexagonal structure. We have tailored the structural parameters to obtain nearly zero flat-top dispersion profile with all normal dispersion profile. The dispersion at 5 μm is -53.13 ps/nm-km with effective mode area of the fundamental mode as 4.18 μm^2. Proposed PCF is suitable for the generation of an ultra-broadband supercontinuum spectrum from 1 μm to 11 μm in MIR domain when pumped with the peak power of 0.75 KW at 5 μm with 50 fs pulse width. Proposed PCF design holds potentiality for applications like early cancer diagnostics, gas sensing and food quality check.

Keywords Photonic crystal fibers · Supercontinuum generation · Nonlinear optics · Photonic integrated circuits

Introduction

In recent years, research in mid-infrared (MIR) supercontinuum has gathered significant attention [1, 2]. Supercconntinuum generation (SCG) is the process where amalgamation of various nonlinear phenomena including stimulated Raman scattering, self-phase modulation (SPM), cross-phase modulation (XPM), and four-wave mixing takes place (FWM). They work together on an intense pump beam, leading to a significant amount of spectral broadening compared to the original pump beam in an optical fiber. This concept of SCG was first introduced in 1970 [3]. It involves interplay between Raman effect, Kerr effect and dispersion properties of the fiber. Among fibers we have a special class of fibers called holey fibres (HFs) or the microstructured fibers (MOFs) or photonic crystal fibers (PCFs) due to their exclusive linear and nonlinear properties they have been a good choice among researchers [4]. The reason for such fascination is their properties such as dispersion management, large mode area with endlessly single-mode operation, higher birefringence, SCG and high power delivery. When compared to traditional optical fibers, PCF requires lesser initial peak power of the input laser pulses. It results in a huge spectral broadening at a very mimimal peak power this fact can be utilized in applications like frequency metrology [5].

The fiber of choice has traditionally been the silica PCF because it can be made endlessly single-mode, its dispersion can be tailored and its zero-dispersion wavelength (ZDW) can move down into the visible region [6]. Chalcogenide glasses have wider transmission range and a greater nonlinear coefficient in a MIR waveband when compared to silica (non-transparent beyond 2.5 μm), sulfur, fluoride and telluride glasses (long wavelength spectrum is limited at 5 μm) [7].The chalcogenide materials are preferred for broadband SCG because of the lower peak power requirement and high nonlinearity. Compared to single glass chalcogenide fibers, step-index fibers with hybrid-glasses offer several benefits. Firstly, they exhibit a higher refractive-index (RI) difference between the fiber core and cladding, enhancing the effectiveness of light confinement in the core and reducing loss. Secondly, the fiber's numerical aperture is increased due to the higher RI difference, facilitating the collection of light into the core during coupling. Lastly, effective controlling over the chromatic dispersion can be achieved by using

✉ Ajeet Kumar
ajeetdph@dtu.ac.in

Drishti Singh Tomer
drishti.tomer@gmail.com

[1] Advanced Photonics Simulation Research Lab, Department of Applied Physics, Delhi Technological University, Delhi 110042, India

Published online: 15 June 2024

multiple glasses. Consequently, a hybrid fiber composed of two types of chalcogenide glass proves to be well-suited for supercontinuum (SC) generation, especially in extending to the M-IR band [8].

Photonic crystals, waveguides, and photonic crystal fibers all leverage the control of light propagation through structural design and material choice. They share foundational principles such as the use of periodic structures and photonic bandgaps, while differing in their specific forms and applications [9–18]. Several theoretical and experimental studies have documented the exploration of silica, tellurite, fluoride, and chalcogenide glasses in the context of step-index fibers, photonic crystal fibers (PCFs), and waveguides for (SCG) [1, 19–29]. The experimental demonstration of the broadening in SC spectrum from 2.1 to 3.1 µm has been performed by Shaw et al. using a PCF in As_2Se_3 chalcogenide glass [19]. The bandwidth of SCG is maximized by Hu et al. using As_2Se_3 chalcogenide PCF for input pulse at the wavelength of 2.5 µm [22]. Petersen et al. reported a Mid-IR SC spectra, spanning 1.4–13.3 µm by using a chalcogenide step-index fiber [9]. To achieve high nonlinearity at 1.55 µm a composite PCF of As_2Se_3 glass and tellurite is modeled by Liao et al. [10]. Saini et al. presented a triangular core graded-index (GI) PCF for SCG [23]. The SCG of bandwidth 1–6 µm has been computationally analyzed by Seifouri et al. in a PCF made up of As_2Se_3-MgF_2 composition [24]. An equiangular spiral shaped PCF structure is theoretically presented by Saini et al. to generate the SC spectrum, spanning 1.2–15 µm [25]. Salem et al. have done the numerical investigation of a tapered As_2S_3 PCF design using a pulse energy of 100 pJ pumping at 4.7 µm for generation of ultra-broadband SC spectrum [26]. Numerical modeling of an As_2Se_3 rib waveguide is reported by Saini et al. for SCG [27]. In our previously published work, we have done numerical modelling of a parabolic core waveguide for SCG [2]. Using a Ge-Sb-S chalcogenide based waveguide, nonlinear characterization is carried out by Choi et al. for the demonstration of the SC spectrum [28]. A Ga-Sb-S based PCF has been studied for coherent SCG by injecting a laser pulse of 50 fs pulse width and 20 kW peak power in 1 cm fiber length at 4.5 µm [29]. The numerical simulations with an experimental extensive study has been performed by Xing et al. to generate SCG spanning, 1.9–2.34 µm at −20 dB power level in a 2 cm long fiber [20]. A 150 mm long $Ge_{11.5}$ As_{24} $Se_{64.5}$ based PCF is numerically simulated for effective SC spanning, 1.4–10 µm using 85 fs input laser pulse at 3.1 µm with peak power of 5 kW [21].Recently, the maximum SC range covering one octave in a chalcogenide step-index fiber with $AsSe_2$ core and As_2S_5 cladding from 1550 to 3300 nm is obtained with 20 cm long fiber at 2000 nm pulses in normal dispersion region [1].

Among the well-known chalcogenides the transparency range of the $AsSe_2$ core glass extended from approximately 0.83 to 18.9 µm, surpassing that of the conventional materials like As_2Se_3 (0.8–10 µm) [30]. In this study, we devised and produced an innovative photonic crystal fiber (PCF) structure by combining $AsSe_2$ and As_2S_5 materials, resulting in a significant difference in refractive indices. The $AsSe_2$-As_2S_5 composition was carefully chosen for their admirable compatibility during the drawing process to prevent interface cracking. Also, the significant contrast between their refractive index facilities efficient control over chromatic dispersion [31].

In our reported design we are experiencing higher supercontinuum broadening at a low input pump power. Also the tradeoff between broadening and coherence has been efficiently addressed. The generated supercontinuum source can be used to discern chemicals, inspect food quality, detect explosives and hazardous gases, ulcer and cancer diagnosis, frequency comb generation, optical imaging, and optical coherence tomography. This paper includes five segments in total. First segment provides the brief introduction and overview of the work. While, the second segment includes about the PCF designing and its importance. In the third segment, provides the extensive illustration of the analytical methodology used. Fourth segment comprises of results and its discussion. At the end, conclusion of the entire work has been provided in section five followed by references.

In this work, we report a mid-IR supercontinuum spectrum extending form 1 –11 µm at 5 µm using a 4 mm long fiber pumped with 50 fs laser pulses of 750 W. In comparison to recently reported result in Table 1 we obtained broader coherent mid-IR supercontinuum spectrum with comparatively lower peak power of input laser pulse.

Table 1 Value of parameters in the design

		Inner cladding region		Outer cladding region
1st helical ring diameter(µm) (d_1)	2nd helical ring diameter(µm) (d_2)	3rd helical ring diameter (µm) (d_3)	δD(µm) ($=d_1-d_2=d_2-d_3$)	Diameter of hexagonal air holes(µm) (d_4)
0.50	0.50	0.50	0.00	0.70
1.10	0.80	0.50	0.30	0.70
1.30	0.90	0.50	0.40	0.70
1.50	1.00	0.50	0.50	0.70

Design principle of PCF

Figure 1(a) shows the cross-sectional view of proposed PCF. The core is structured using $AsSe_2$ with diameter d_{core}. Cladding is hybrid with two regions, inner cladding and outer cladding. Inner cladding is a graded helical structure around a fixed core. Outer cladding is a hexagonal structure. As_2S_5 is used as a cladding material in combination with air holes. The cladding comprises of 6 layers of hexagonal structured air-holes, where the first 3 layers of air holes have been eliminated from the center to design inner graded helical structure. The inner cladding structure consists of 8 helical patterns and each pattern comprises 3 air holes that are positioned at $r_k = r_{k-1} + 1.5*d$ from the center. The angular variation is given as $\theta_k = (k * 360^\circ) / (2 * K)$ and K reflects the number of spherical rings (K = 3 in our case) as shown in Fig. 1(c) &1(d). For effective detention of light and its propagation through the core forked shift of air holes along with grading is presumed. First three layers of the proposed graded – index (GI) helical shape are arranged with diameter of 1st, 2nd and 3rd ring as $d_1 > d_2 > d_3$,. The value of d_3 is constant as 0.50 µm but d_1 and d_2 has been changed. The distance from the center of the first ring is constant and represented as r_0, which is 2.0 µm. The pitch (Λ= spacing between two air holes) in other rings is optimized. Λ is changed from 1.8 µm to 2.2 µm with a period of 0.2 µm. Grading has been done in a systematic manner where 1st ring diameter is varied as (d_1 = 1.10 µm, 1.30 µm, 1.50 µm), 2nd ring diameter (d_2 = 0.80 µm, 0.90 µm, 1.00 µm), 3rd ring diameter (d_3 = 0.50 µm), core diameter is also varied as (d_{core} = 1.2 µm, 1.4 µm, 1.6 µm, 1.8 µm). Amount of successive grading ($\delta D = d_1 - d_2 = d_2 - d_3$ = 0.30 µm, 0.40 µm, 0.50 µm). The next three rings which forms the outer cladding region have a hexagonal shape of constant diameter (d_4 = 0.70 µm). The distance of first hexagonal ring from the center is 5.6 µm. The distance between the consecutive hexagonal rings is kept constant which is 1 µm. The value of structural parameters have been shown in Table 1 for better understanding. The cylindrical PML is applied at the periphery of the cladding for eradicating the reflection effect occurring at the boundary. Here, the thickness of perfectly matched layer is fixed [23, 32, 33].

Achieving a profile with all-normal dispersion (ANDi) as well as nearly zero dispersion is the purpose of fabricating such a structure. Since, it helps in achieving broader SCG and low noise. Simulated electric field distribution of traversing mode in proposed PCF design has been illustrated in Fig. 1(b). From the fabrication point of view, the reported PCF structure can be fabricated using the well-known rod in tube fabrication technique [34].

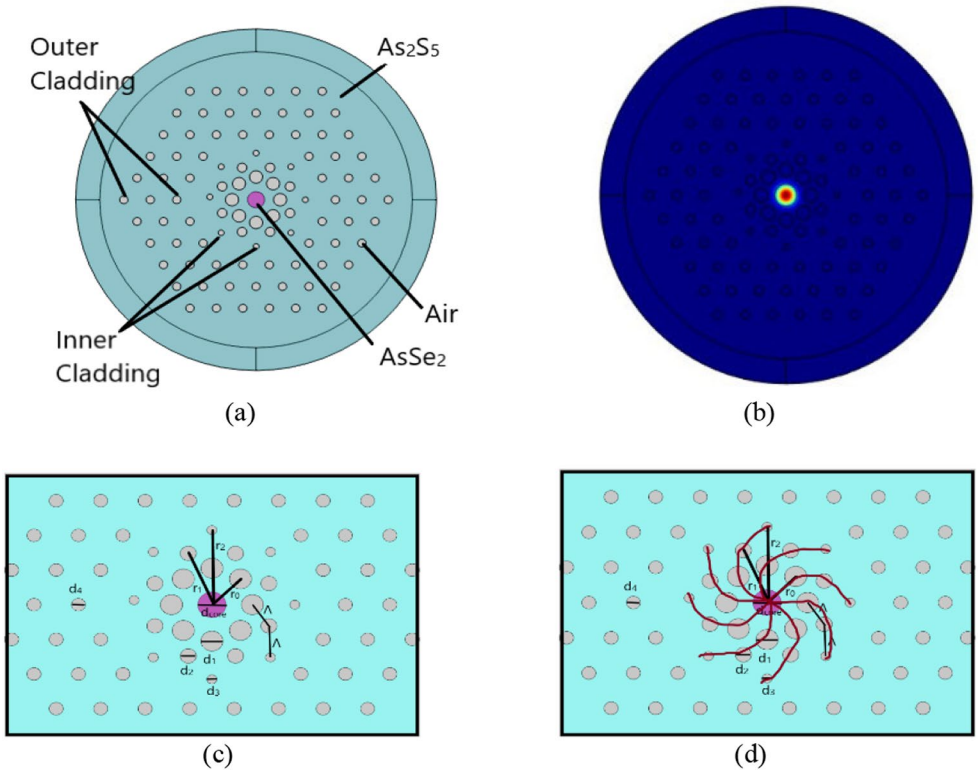


Fig. 1 (*a*) The cross-sectional perspective of the suggested PCF in a transverse view; (*b*) the distribution of electric field for the fundamental mode at 5 µm ; (*c*) structural parameters of the PCF ; and (*d*) displaying spiral nature of the inner cladding for better understanding

Analytical methodology

Linear characteristics

The PCF structure and the effective mode index (n_{eff}) of the fundamental mode is determined using COMSOL software which uses full-vectorial finite element (FEM) method. The wavelength dependent refractive index of $AsSe_2$ chalcogenide glass and As_2S_5 materials was calculated by using the Sellmeier equation and the constant values for $AsSe_2$ ($B_1=6.751$, $B_2=0$, $B_3=0.1768$, b_1^2 0.1034 μm^2, b_2^2 = 153.7990 μm^2, b_3^2 = 10^{-32} μm^2) and As_2S_5 ($B_1=2.1361$, $B_2=0.0693$, $B_3=1.7637$, $b_1^2=0.0954$ μm^2, $b_2^2=225.0001$ μm^2, $b_3^2=2.1783\times10^{-7}$ μm^2) are known [35, 36]

$$n^2(\lambda) = 1 + \sum_{k=1}^{N} \frac{B_k \lambda^2}{\lambda^2 - b_k^2} \tag{1}$$

For Arsenic selenides the wavelength range for sellmeier equation to be valid is typically 1 –14 μm and in some cases it has been further extended up to 0.6 –15 μm [37].

The comprehensive dispersion profile of the intended PCF framework is assessed by employing the following relationship

$$D(\lambda) = -\frac{\lambda}{c}\frac{\partial^2 Re(n_{eff})}{\partial \lambda^2} \tag{2}$$

Re (n_{eff}) denotes the real component of the effective mode index and c is the speed of light in free space. Using the following equation, the effective mode area (A_{eff}) of the propagating mode has been conducted [38]

$$A_{eff} = \frac{\left(\iint_{-\infty}^{\infty} |E|^2 dxdy\right)^2}{\left(\iint_{-\infty}^{\infty} |E|^4 dxdy\right)} \tag{3}$$

Where, the amplitude of the electric field of the light is represented by E.

3.2. Nonlinear characteristics

In the simulations, the initial input laser pulse considered is a hyperbolic secant, as depicted in the equation below.

$$\mathrm{A}(0,\mathrm{T}) = \sqrt{\mathrm{P}_0}\mathrm{sech}\frac{\mathrm{T}}{\mathrm{T}_0} \tag{4}$$

The equation introduces various components: The pulse envelope A (0, T), characterizes the shape of the pulse. P_0, the peak power of the input pulse. T_0 is full-width-at-half-maxima (FWHM) divided by 1.7627, specifically tailored for hyperbolic secant pulses. T is the concurrent frame associated with the group velocity (GV) of the pulse envelope.

The following generalized nonlinear Schrödinger equation is solved using Split-Step Fourier method to obtain the widening of the spectral range observed in the generated SC within the PCF structure [39]

$$\frac{\partial \tilde{A}'}{\partial z} = i\tilde{\gamma}(\omega)\exp\left(-\hat{L}(\omega)z\right) \mathcal{F}\left\{\bar{A}(z,T)\int_{-\infty}^{\infty} R(T')\left|\bar{A}(z,T-T')\right|^2 dT'\right\} \tag{5}$$

Where, $\tilde{A}'$ is envelope or the covering of the output pulse which represents its modulation amplitude or shape, $\bar{A}(z,T)$, the envelope at distance z at time t. $\bar{A}(z,\omega)$, the Fourier transform of $\bar{A}(z,T)$. The frequency-dependent nonlinear coefficient $\bar{\gamma}(\omega)$ is obtained by taking into account its dependency on the frequency

$$\bar{\gamma}(\omega) = \frac{n_2 n_0 \omega}{c n_{eff}(\omega) A_{eff}^{1/4}(\omega)} \tag{6}$$

n_2 corresponds to the nonlinear refractive index, which is specifically 2.3×10^{-17} m^2/W for $AsSe_2$ chalcogenide glass [40]. The symbol n_0 represents the glass refractive index at the wavelength needed to calculate n_2 and c being the speed of light in free space (vacuum). The expression for the altered variables are stated as

$$\tilde{A}'(z,\omega) = \tilde{A}(z,\omega)\exp(-\hat{L}(\omega)z) \tag{7}$$

The following equation shows the linear operator $\hat{L}(\omega)$as

$$\hat{L}(w) = i(\beta(\omega) - \beta(\omega_0) - \beta_1(\omega_0)[\omega-\omega_0]) - \frac{\alpha(\omega)}{2} \tag{8}$$

β, propagation constant; β_1, the inverse of the group velocity of pulse envelope; α, the total loss and; ω_0, the reference frequency, within the PCF design. The Raman response function$R(t)$ is defined as

$$R(t) = (1-f_R)\delta(T) + f_R\frac{\tau_1^2+\tau_2^2}{\tau_1\tau_2^2}\exp\left(-\frac{T}{\tau_2}\right)\sin\left(\frac{T}{\tau_1}\right)\mathrm{H}(\mathrm{T}) \tag{9}$$

Here, H (T) being the Heaviside step function depends on T. For the $AsSe_2$ glass, Raman parameters f_R,τ_1,τ_2 are known [35].

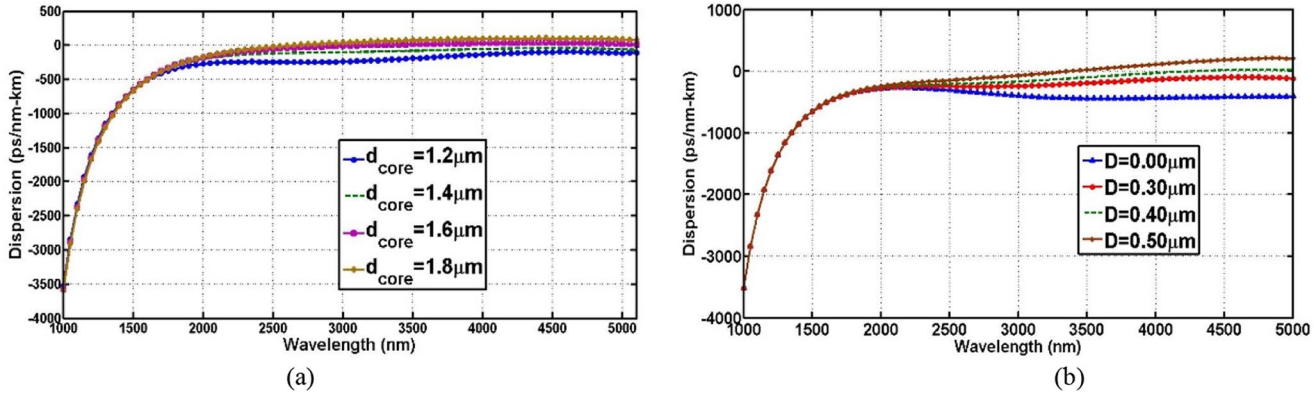


Fig. 2 (**a**) Effect of core Width "d_{core}" on dispersion characteristic of the PCF with $d_1 = 1.10$ μm, $d_2 = 0.80$ μm, $\delta D = 0.30$ μm and $\Lambda = 2.2$ μm ; (**b**) Effect of grading on the dispersion characteristic of the PCF for $d_{core} = 1.6$ μm and $\Lambda = 2.2$ μm

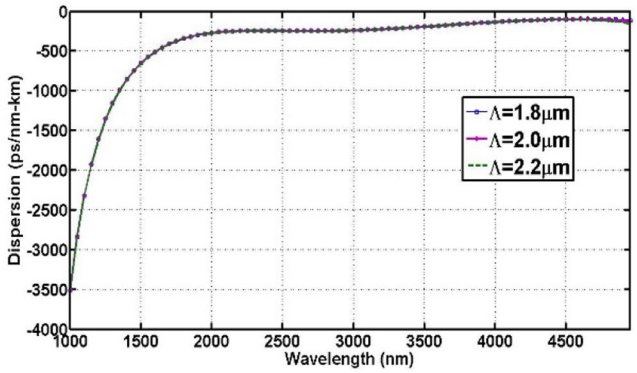


Fig. 3 Effect of pitch on the dispersion characteristic of the PCF for $d_{core} = 1.6$ μm and $\delta D = 0.30$ μm

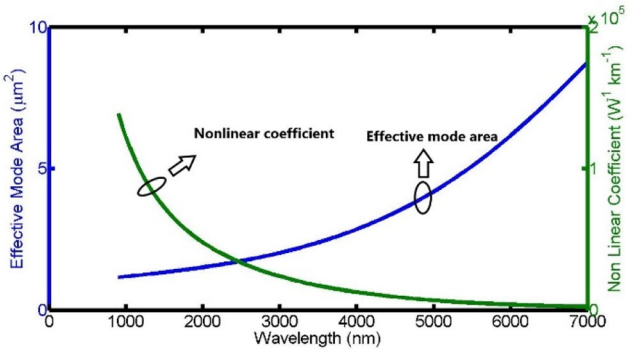


Fig. 4 The spectral changes in the effective mode area and nonlinear coefficient are depicted with respect to wavelength for PCF

Results and discussion

For a PCF structure the broadening in the supercontinuum spectrum primarily depend upon the nonlinearity of material and the dispersion characteristics [13].We have optimized the core diameter of the proposed design for d_{core} (1.2 μm, 1.4 μm, 1.6 μm and 1.8 μm) as shown in Fig. 2(a). Dispersion profile shows all normal dispersion behavior for all d_{core} except for 1.8 μm. Soliton fission process along with the emission of the dispersive waves is responsible for the broadening in the supercontinuum spectrum if pumped in the anomalous dispersion region. The SCG by the process of soliton fission is vulnerable to the laser shot noise which makes supercontinuum unstable.

As depicted in Fig. 2 (a) with 1.6 μm as the core diameter, $d_1 = 1.10$ μm, $d_2 = 0.80$ μm, $\delta D = 0.30$ μm and pitch = 2.2 μm the profile is more flat and all normal with minimum value of dispersion around 4.8 μm as compared to rest of the core diameters. So, we have opted for 1.6 μm as a core diameter. Further, the air holes in the inner cladding region has been graded in the decreasing order of diameter. Figure 2(b) shows dispersion for δD (amount of grading). Among the evaluated δD (0.00 μm, 0.30 μm, 0.40 μm and 0.50 μm) it has been found that both $\delta D = 0.00$ μm(non-graded) & $\delta D = 0.30$ μm is showing all normal dispersion profile whereas others have both normal and anomalous dispersion regime. But between these two, we have considered $\delta D = 0.30$ μm, along with being all normal it is having less dispersion. So it will result in more broadening as well as good coherency. Figure 3 shows that on changing the pitch (Λ) of the air holes in the inner cladding region there is no significant change in the nature of the profile. On changing pitch the effective refractive index of the profile has not changed much which could possibly affect dispersion curve, since it is same so the profile is also same for different pitch values. Considering the availability of a commercial available laser source we have taken 5 μm as a pump wavelength. At 5 μm the dispersion is -53.13 ps/nm-km for 1.6 μm core diameter, $\delta D = 0.30$ μm and $\Lambda = 2.2$ μm. In Fig. 4, effective mode area vs. nonlinearity shows the obvious trend when plotted against wavelength. At 5 μm effective mode index, effective mode area and nonlinearity has been evaluated and found to be 2.278, 4.18 μm^2 and 6900 $W^{-1}Km^{-1}$ respectively.

The complete loss is sum of both material and confinement loss in a fiber. The value of confinement loss is less in comparison to the material loss of the $AsSe_2$ chalcogenide glass, which shows absorption peak in the loss curve around 2.9 μm due to the OH impurity in the chalcogenide glass, and other absorption peak around 12.7 μm because of the Se-OH bonds [41]. The confinement loss was very less in comparison to the material loss and its value is 1.19×10^{-2} dB/m at 5 μm.

SCG has been evaluated at pump wavelength of 5 μm for various values of fiber length, pulse width and peak power with 1 kHz repetition rate. The maximum broadening is observed for 4 mm fiber length, 750 W peak power and 50 fs pulse width. Figure 5 shows variation of intensity with length from 2 mm to 8 mm, it is seen that the broadening is very less below 4 mm and above 4 mm noise is increasing. Similarly, in Fig. 6 the length variation in a spectrum from 4 mm to 8 mm depicts that on increasing the length at-40 dB power level the spectrum is experiencing a shift towards longer wavelength region. In the beginning the broadening has occurred due to SPM which is later dominated by (OWB) optical wave breaking. Further increase in length add undesirable fluctuations in the supercontinuum spectrum and deteriorate coherence. So we have considered 4 mm as optimum fiber length. Also, from Fig. 7 it is evident that on increasing length the coherency is also getting disturbed.

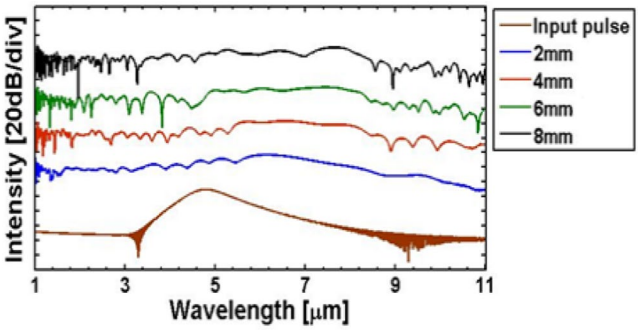


Fig. 5 The advancement of spectral SC broadening along the length of the PCF, induced by 50 fs laser pulses with 750 W pump power at 5.0 μm

In Fig. 8, as the pump power is incremented from 750 W to 2750 W the spectrum gets shifted towards the longer wavelength region. As the pumping power was further

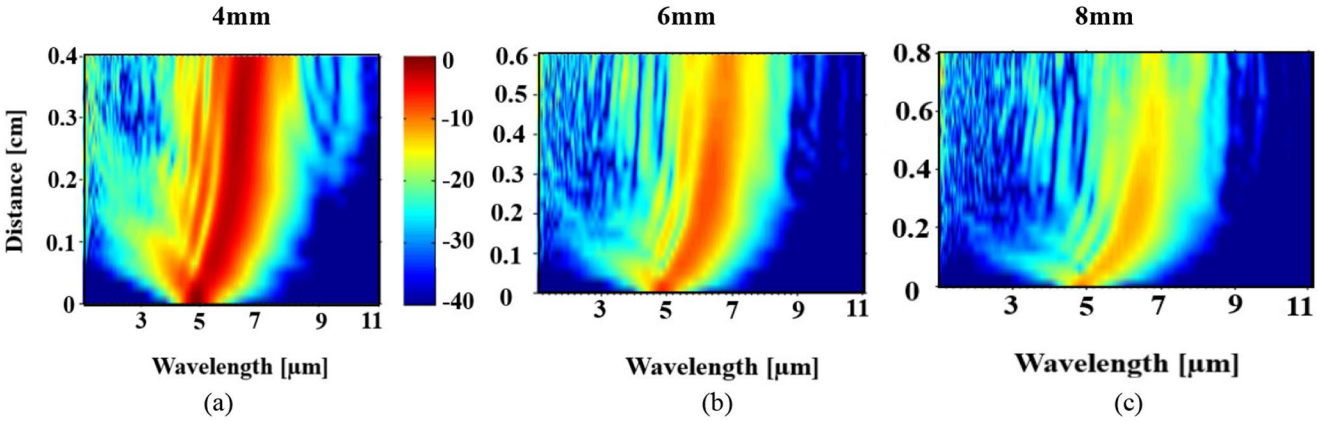


Fig. 6 The advancement of spectral SC broadening along the length (**a**) 4 mm; (**b**) 6 mm; (**c**) 8 mm of the PCF, induced by 50 fs laser pulses with 750 W pump power at 5.0 μm

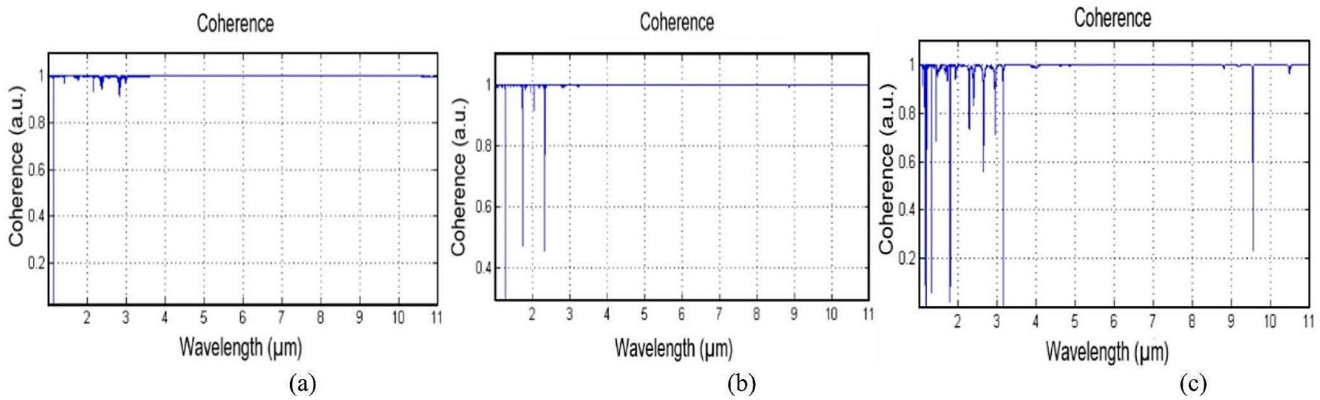


Fig. 7 The changes in the coherence evolution at 5 micron in response to the input pulse peak power 750 W,50 fs pulse width along the length of the fiber (**a**) 4 mm; (**b**) 6 mm; (**c**) 8 mm

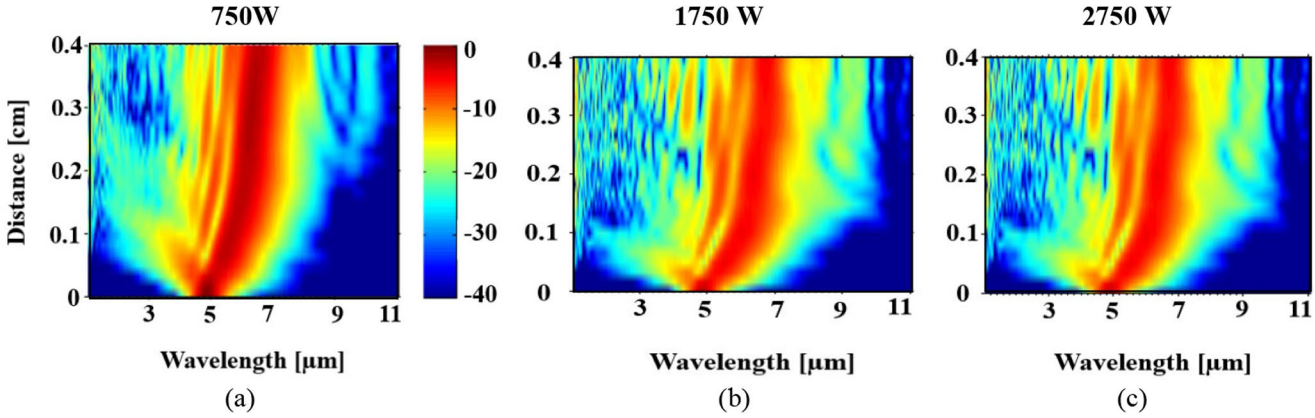


Fig. 8 The advancement of spectral SC broadening with the pump power, (**a**) 750 W; (**b**) 1750 W; (**c**) 2750 W; induced by 50 fs laser pulses with 4 mm fiber length at 5.0 μm

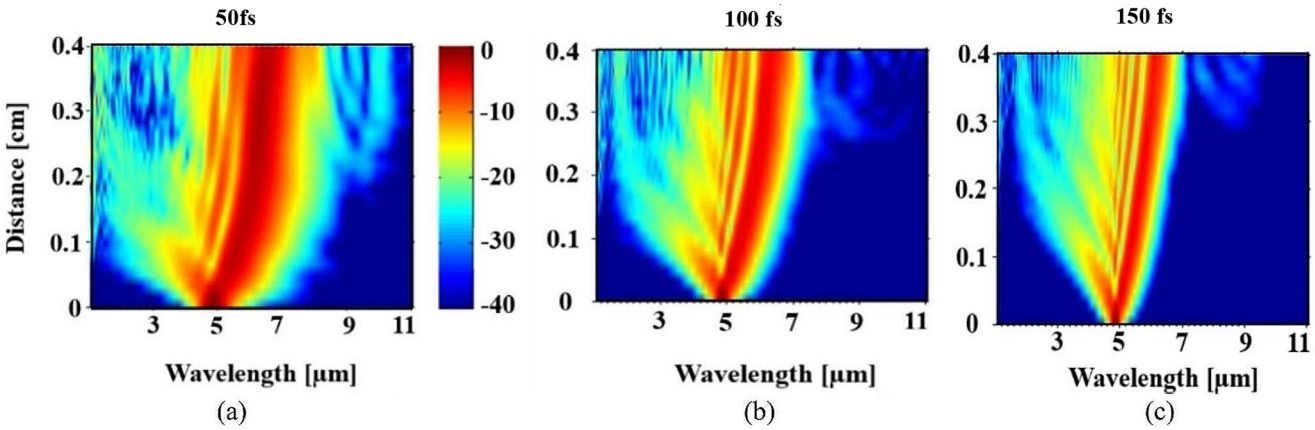


Fig. 9 The advancement of spectral SC broadening at 5 μm induced by (**a**) 50 fs; (**b**) 100 fs; (**c**) 150 fs laser pulses at 4 mm fiber length ,750 W pump power

increased the SC started to degrade due to absorption at the coating [42] In Fig. 9, it can be observed as the value of T_{FWHM} is increased from 50 femtosecond to 150 femtosecond the resultant spectra began to get (thinner) narrow down. This is due to the Self phase modulation effect. It is worth noting that shorter pulses tend to get broader SC spectra. Hence, 50 fs pulses are giving stable spectrum.

For 50 fs laser pulse at 5 μm, β_2 is 1.15×10^{-25} ; ($T_0 = T_{FWHM}/1.763$) ; the nonlinear length ($L_{NL} = (\gamma P_0)^{-1}$) and dispersion length ($L_D = (\beta_2/T_0^{\,2})^{-1}$) for proposed PCF are 1.9×10^{-4} m and 2.17×10^{-2} m respectively. In order to achieve a broader supercontinuum generation, it is desirable to have a nonlinear length that is comparable to or shorter than the dispersion length. This ensures that nonlinear effects play a significant role in shaping the spectrum of the input pulse, while dispersion does not excessively broaden the pulse before nonlinear effects can take place. The soliton order, $N \approx L_D/L_{NL} \approx 10$ and the soliton fission length, $L_{fiss} = L_D/N \approx 2.17$ mm is evaluated using these relations. In the starting phase of pulse propagation, self -phase modulation (SPM) is responsible for spectral broadening whereas at longer wavelength optical wave breaking is responsible for broadening.

Within the range of 1.3–1.7 μm, it is advantageous to modify the probing wavelength in order to enhance the imaging capabilities of optical coherence tomography (OCT) systems for biological tissue [41]. Additionally, near-to-mid-IR supercontinuum pulses that are repeatable from shot to shot are crucial for enhancing image penetration depth and regaining a 10 dB signal-to-noise ratio [43]. In this paper, we concentrate on the extensive range of SC spectra achieved with minimal pump power within a short length of the suggested PCF structure. Further, to explore the degree of coherence (DoC) of the obtained spectra, we have calculated coherence in Fig. 7 for length 4 mm, 6 mm and 8 mm respectively. The first-order DoC is given as

$$|g_{12}(\lambda(t_1 - t_2)| = \frac{|< E_1^*(\lambda, t_1)E_2(\lambda, t_2) > |}{(|< E_1(\lambda, t_1)^2 < \left|E_2(\lambda, t_2)^2\right| >)^{1/2}} \tag{10}$$

Table 2 Comparison against previously documented outcomes of SC spectrum in chalcogenide PCFs

References	Chalcogenide type	Pump Wavelength (μm)	Fiber length	Peak power (kW)	SCG bandwidth (μm)
[44]	$As_{38.8}$ $Se_{61.2}$ PCF	3.7	5 cm	0.88	2.9–4.575
[45]	As_2S_5 – Borosillicate PCF	2.5	4 mm	28.16	1–5
[46]	As_2Se_3 PCF	2.5	10 mm	20	1–10
[29]	$Ga_8Sb_{32}S_{60}$ PCF	4.5	1 cm	20	1.65–9.24
[42]	$GeSe_2$-As_2Se_3-PbSe PCF	3.1	10 mm	0.95	1–11
Proposed work	$AsSe_2$- As_2S_5 PCF	5.0	4 mm	0.75	1–11

Electric fields of independent SCG pairs are given by $E_1(\lambda, t_1)$ and $E_2(\lambda, t_2)$. These electric fields pertains to immediately consecutive generated SCG spectra with the angular brackets indicating an average over the entire pulse ensemble. To center on the wavelength dependence of the coherence, $\Delta t = t_1 - t_2 = 0$ was adopted.

The DoC function is bounded as $0 \leq |g_{12}(\lambda, t_1 - t_2)| \leq 1$. The value 0 signifies completely incoherent spectrum while the maximal value 1 represents perfect spectral coherence. We have computed the DoC function especially to examine the wavelength dependence of the coherence [42].

Finally we conclude that our proposed structure is able to generate broader and comparable SCG bandwidth with lower pump power in comparison to previously reported work, mentioned in Table 2.

As can be inferred from the table $As_{38.8}$ $Se_{61.2}$ PCF [44] and As_2S_5 – Borosillicate PCF [45] has a limited broadening up to 5 μm, As_2Se_3 PCF [46] and $Ga_8Sb_{32}S_{60}$ PCF [21] has broadening around 1 to 10 μm but the peak power used is quite high around 20 kW. In $GeSe_2$-As_2Se_3-PbSe PCF [42] the broadening is comparable but the peak power used is higher as 0.95 kW with 10 mm fiber length as compared to 0.75 kW and 4 mm fiber length in our case.

Since the focus of this work is to produce greater broadening with low peak power as it offers several advantages in SCG. It can reduce nonlinear effects which can distort the spectrum and limit the bandwidth of the supercontinuum.It leads to better performance in applications where spectral stability is crucial, such as optical coherence tomography (OCT) and spectroscopy. With lower peak power, the supercontinuum spectrum tends to be more stable over time. It enhances Signal-to-Noise Ratio which improves the quality of measurements and imaging in various applications. It reduces photodamage. Also, it often corresponds to longer pulse durations, which can be beneficial for certain applications that require longer interaction times with the sample, such as material processing and biomedical imaging.

Overall, using lower peak power in supercontinuum generation can result in more stable, high-quality spectra with reduced nonlinear effects and improved safety for sensitive samples.

Conclusion

In summary, the especially designed graded index hybrid PCF structure with $AsSe_2$- As_2S_5 offers all normal dispersion (ANDi) at pump wavelength of 5 μm. The essence of this work lies in the generation of mid-IR supercontinuum an efficient broadband coherent mid-IR SC spectrum extending from 1 to 11 μm is reported in a 4 mm long PCF when subjected to 750 W, 50 fs laser pulse at a wavelength of 5 μm. From an application standpoint, the utilization of a PCF-based broadband coherent NIR to MIR SC light holds great significance in various critical domains, including early cancer diagnostics, food quality control, security, sensing, and the recognition of a sustained optical frequency comb for enabling high-precision spectroscopy and optical coherence tomography. Additionally, the reported PCF, which generates NIR to MIR SC light, has the potential to significantly expand the range of functionalities achievable in future photonic integrated circuits.

Acknowledgements Drishti Singh Tomer highly acknowledges the support from the Delhi Technological University, Delhi for providing the necessary support to conduct this work.

Data availability No data was used for the research described in the article.

References

1. Y. Qu, W. Sun, J. Cao, H. Chen, H. Jia, Coherent ultrabroad orbital angular momentum supercontinuum generation in an $AsSe_2$ -As_2S_5 microstructured fiber with all-normal dispersion. Opt. Commun. **497**, 127191 (2021). https://doi.org/10.1016/j.optcom.2021.127191
2. D.S. Tomer, Design and numerical modeling of chalcogenide parabolic - core waveguide for on - chip supercontinuum generation extending from near - IR region to mid - IR region, 1–7, (2024) https://doi.org/10.1002/mop.34105
3. R.R. Alfano, S.L. Shapiro, Emission in the region 4000 to 7000 Å Via Four-Photon Coupling in Glass. Phys. Rev. Lett. **24**, 584–587 (1970). https://doi.org/10.1103/PhysRevLett.24.584
4. P. Chauhan, A. Kumar, Y. Kalra, Optik A dispersion engineered silica-based photonic crystal fi ber for supercontinuum generation in near-infrared wavelength region. Opt. - Int. J. Light Electron. Opt. **187**, 230–237 (2019). https://doi.org/10.1016/j.ijleo.2019.03.107
5. T.S. Saini, A. Kumar, R.K. Sinha, Broadband mid-infrared Supercontinuum Spectra spanning 2–15 μm using As_2Se_3 Chalcogenide

Glass Triangular-Core graded-index Photonic Crystal Fiber. J. Light Technol. **33**, 3914–3920 (2015). https://doi.org/10.1109/JLT.2015.2418993
6. J.K. Ranka, R.S. Windeler, A.J. Stentz, Visible continuum generation in air–silica microstructure optical fibers with anomalous dispersion at 800 nm. Opt. Lett. **25**, 25 (2000). https://doi.org/10.1364/ol.25.000025
7. W. Yuan, 2–10 µm mid-infrared supercontinuum generation in as 2 Se 3 photonic crystal fiber. Laser Phys. Lett. **10**, 095107 (2013). https://doi.org/10.1088/1612-2011/10/9/095107
8. Weiqing, Gao, Supercontinuum generation in a step-index chalcogenide fiber with $AsSe_2$ core and As_2S_5 cladding. Jpn J. Appl. Phys. **55** (2016). https://doi.org/10.7567/JJAP.55.122201
9. S. Ankita, B. Bissa, A. Suthar, Bhargava, Study transmission characteristics of graded photonic crystal as low pass filter. J. Phys. : Conf. Ser. **2603**, 012046 (2023). https://doi.org/10.1088/1742-6596/2603/1/012046
10. S. Ankita, B. Bissa, A. Suthar, Bhargava, High pass filter based on one-dimensional graded photonic crystal high pass filter based on one-dimensional graded photonic. J. Phys. : Conf. Ser. **2335**, 012006 (2022). https://doi.org/10.1088/1742-6596/2335/1/012006
11. S. Bissa, B. Suthar, C. Nayak, A. Bhargava, Results in Optics An Improved Optical Biosensor design using defect / metal multilayer photonic crystal for Malaria diagnosis. Results Opt. **9**, 100304 (2022). https://doi.org/10.1016/j.rio.2022.100304
12. M. Sharma, F. Amirkhan, S.K. Mishra, D. Sengupta, Y. Messaddeq, F. Blanchard, B. Ung, Annular core photonic crystal fiber for transmission of endlessly mono-radial vortex beams, *in* OSA Advanced Photonics Congress (AP) (IPR, NP, NOMA, Networks, PVLED, PSC, SPPCom,SOF), L. Caspani, A. Tauke-Pedretti, F. Leo, and B. Yang, eds., OSA Technical Digest (Optica Publishing Group, 2020), paper SoM3H.7, (2020) https://doi.org/10.1364/SOF.2020.SoM3H.7
13. J.M. Dudley, G. Genty, S. Coen, Supercontinuum generation in photonic crystal fiber. Rev. Mod. Phys. **78**, 1135–1184 (2006). https://doi.org/10.1103/RevModPhys.78.1135
14. M. Sharma, F. Amirkhan, S.K. Mishra, D. Sengupta, Y. Messaddeq, F. Blanchard, B. Ung, Transmission of Orbital Angular Momentum and Cylindrical Vector beams in a large-bandwidth annular core Photonic Crystal Fiber. Fibers. **8**(4), 22 (2020). https://doi.org/10.3390/fib8040022
15. B. Suthar, A. Bhargava, Enhanced optical sensor for waterborne bacteria-based photonic crystal using graded thickness index. Appl. Nanosci. **13**, 5399–5406 (2023). https://doi.org/10.1007/s13204-023-02809-2
16. S. Ankita, B. Bissa, A. Suthar, Bhargava, Graded photonic crystal as improved sensor for nanophotonic application. Macromol.Symp. **401**, 2100319 (2022). https://doi.org/10.1002/masy.202100319
17. S. Ankita, B. Bissa, A. Suthar, Bhargava, Optical sensor for biomedical application based on photonic crystal with double defect. Mater. Today: Proc. **62**(Part 8), 5407–5410 (2022). ISSN 2214–7853. https://doi.org/10.1016/j.matpr.2022.03.616
18. A. Kumar, T.S. Saini, R.K. Sinha, Design and analysis of photonic crystal biperiodic waveguide structure based optofluidic-gas sensor. Optik. **126**(24), 5172–5175 (2015)
19. L.B. Shaw, P.A. Thielen, F.H. Kung, V.Q. Nguyen, J.S. Sanghera, IR Supercontinuum Generation in As-Se Photonic Crystal Fiber (2005), pp. 10–12. https://doi.org/10.1364/ASSP.2005.TuC5
20. E. Polytechnique, F. De Lausanne, Linearly chirped mid-infrared supercontinuum in all-normal-dispersion chalcogenide photonic crystal fibers. Optic Exp. **26** 2423–2428 (2018) https://doi.org/10.1364/OE.26.019627
21. S. Vyas, T. Tanabe, G. Singh, M. Tiwari, Broadband supercontinuum generation and Raman response in $Ge_{11.5}As_{24}Se_{64.5}$ based chalcogenide photonic crystal fiber, in *International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)* (IEEE, 2016), pp. 607–611. https://doi.org/10.1109/ICCTICT.2016.7514651
22. H. Jonathan, R. Curtis, L. Menyuk, J.S. Brandon Shaw, Sanghera, I.D. Aggarwal, Maximizing the bandwidth of supercontinuum generation in As_2Se_3 chalcogenide fibers. Opt. Express. **18**, 6722–6739 (2010). https://doi.org/10.1364/OE.18.006722
23. T.S. Saini, A. Kumar, R.K. Sinha, Broadband mid-infrared Supercontinuum Spectra spanning 2–15 µm using As_2Se_3 Chalcogenide Glass Triangular-Core graded-index Photonic Crystal Fiber. J. Lightwave Technol. **33**, 3914–3920 (2015). 18, 15 Sept.15. https://doi.org/10.1109/JLT.2015.2418993
24. M. Seifouri, M.R. Alizadeh, Supercontinuum generation in a highly nonlinear chalcogenide/ MgF2 hybrid photonic crystal fiber. IJOP. **12**(1), 69–78 (2018). http://ijop.ir/article-1-277-en.html
25. T.S. Saini, A. Baili, A. Kumar, R. Cherif, Design and analysis of equiangular spiral photonic crystal fiber for mid-infrared supercontinuum generation, (2015). https://doi.org/10.1080/09500340.2015.1051600
26. A. Ben Salem, R. Cherif, M. Zghal, Tapered As_2Se_3 chalcogenide photonic crystal fiber for broadband mid-infrared supercontinuum generation. Optic Eng. **1**, 9–10 (2011).https://doi.org/10.1364/FIO.2011.FMG6
27. T.S. Saini, A. Kumar, R.K. Sinha, Design and modelling of dispersion-engineered rib waveguide for ultra broadband mid-infrared supercontinuum generation. J. Mod. Opt. **64**, 143–149 (2017). https://doi.org/10.1080/09500340.2016.1216190
28. J.W. Choi, Z. Han, B. Sohn, G.F.R. Chen, C. Smith, L.C. Kimerling, K.A. Richardson, A.M. Agarwal, D.T.H. Tan, Nonlinear characterization of GeSbS chalcogenide glass waveguides. Nat. Publ Gr. 1–8 (2016). https://doi.org/10.1038/srep39234
29. A. Medjouri, D. Abed, Z. Becer, Opto-Electronics Review Numerical investigation of a broadband coherent supercontinuum generation in Ga_8 Sb_{32} S_{60} chalcogenide photonic crystal fiber with all-normal dispersion. Opto-Electronics Rev. **27**, 1–9 (2019). https://doi.org/10.1016/j.opelre.2019.01.003
30. T. Cheng, Y. Kanou, X. Xue, D. Deng, M. Matsumoto, T. Misumi, T. Suzuki, Y. Ohishi, Mid-infrared supercontinuum generation in a novel $AsSe_2$ -As_2S_5 hybrid microstructured optical fiber. Optic. Exp. **22** 23019–23025 (2014). https://doi.org/10.1364/OE.22.023019
31. Numerical study on supercontinuum, Generation by different optical modes in $AsSe_2$ -As_2S_5 chalcogenide microstructured fiber, **57** 382–390, (2018) https://doi.org/10.1364/AO.57.000382
32. S. Chandru, T. Shankar, A. Rajesh, Investigation on nanocore-centred nonlinear PCF for high nonlinearity and low dispersion. Ceram. Int. **48**, 16042–16048 (2022). https://doi.org/10.1016/j.ceramint.2022.02.148
33. A.G. Chaitanya, T.S. Saini, A. Kumar, R.K. Sinha, Ultra broadband mid-IR supercontinuum generation in $Ge_{11.5}As_{24}Se_{64.5}$ based chalcogenide graded-index photonic crystal fiber: design and analysis. Appl Opt. **55**(36), 10138–10145 (2016). https://doi.org/10.1364/AO.55.010138. PMID: 28059256
34. T. Cheng, T.H. Tuan, L. Liu, X. Xue, M. Matsumoto, H. Tezuka, T. Suzuki, Y. Ohishi, Mid-infrared Dispersive Waves Generation in an All-solid $AsSe_2$-As_2S_5 Microstructured Optical Fiber, in Frontiers in Optics, OSA Technical Digest (online) (Optica Publishing Group, 2016), paper JTh2A.38 (2016). https://doi.org/10.1364/FIO.2016.JTh2A.38
35. M. Diouf, A. Wague, M. Zghal, Numerical investigation of anultra-broadband coherent mid-infrared supercontinuum in a chalcogenide$AsSe_2$-As_2S_5 multimaterial photonic crystal fiber. J. Opt. Soc. Amer B **36**(2), 8–14 (2019). https://doi.org/10.1364/JOSAB.36.0000A8

36. T. Cheng, Y. Kanou, D. Deng, X. Xue, M. Matsumoto, T. Misumi, T. Suzuki, Y. Ohishi, Fabrication and characterization of a hybrid fiber with a large refractive index difference. Optic. Exp. **22** 13322–13329 (2014). https://doi.org/10.1364/OE.22.013322
37. J.S. Sanghera, L.B. Shaw, I.D. Aggarwal, Chalcogenide Glass-Fiber-based Mid-IR sources and applications. IEEE J Sel. Top. Quant. Electron. **15**, 114–119 (2009). https://doi.org/10.1109/JSTQE.2008.2010245
38. G.P. Agrawal, Nonlinear Fiber Optics, in: *Nonlinear Sci. Daw. 21st Century* (Springer Berlin Heidelberg, Berlin, Heidelberg, Year), pp. 195–211. https://doi.org/10.1007/3-540-46629-0_9
39. M. Dudley, J.R. Taylor (eds.), *Supercontinuum Generation in Optical Fibers* (Cambridge University Press, 2010). https://doi.org/10.1017/CBO9780511750465
40. L. Liu, T. Cheng, K. Nagasaka, H. Tong, G. Qin, T. Suzuki, Y. Ohishi, Coherent mid-infrared supercontinuum generation in all-solid chalcogenide microstructured fibers with all-normal dispersion. Opt. Lett. **41**, 392 (2016). https://doi.org/10.1364/OL.41.000392
41. U. Sharma, E.W. Chang, S.H. Yun, Long-wavelength optical coherence tomography at 17 µm for enhanced imaging depth. Opt. Express. **16**, 19712 (2008). https://doi.org/10.1364/OE.16.019712
42. P. Chauhan, A. Kumar, Y. Kalra, Optical Fiber Technology Numerical exploration of coherent supercontinuum generation in multicomponent $GeSe_2$ -As_2Se_3 -PbSe chalcogenide based photonic crystal fi ber. Opt. Fiber Technol. **54**, 102100 (2020). https://doi.org/10.1016/j.yofte.2019.102100
43. H. Kawagoe, S. Ishida, M. Aramaki, Y. Sakakibara, E. Omoda, H. Kataura, N. Nishizawa, Development of a high power supercontinuum source in the 17 µm wavelength region for highly penetrative ultrahigh-resolution optical coherence tomography. Biomed. Opt. Express. **5**, 932 (2014). https://doi.org/10.1364/BOE.5.000932
44. Flat coherent supercontinuum, Source in $As_{38.8}$ $Se_{6.12}$ chalcogenidephotonic crystal fiber with all-normal dispersion engineering at avery low input energy. Appl. Opt. **56**(2), 163–169 (2017). https://doi.org/10.1364/AO.56.000163
45. A. Ben Salem, Ultraflat-top midinfrared coherent broadband supercontinuum using all normal As_2S_5 -borosilicate hybrid photonic crystal fiber. Opt. Eng. **55**, 5–12 (2016). https://doi.org/10.1117/1.OE.55.6.066109
46. M. Kalantari, A. Karimkhani, H. Saghaei, S.C. Opt, - Int. J. Light Electron. Opt. (2017). https://doi.org/10.1016/j.ijleo.2017.12.014

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Applied Intelligence
https://doi.org/10.1007/s10489-024-05598-5

A lightweight convolutional swin transformer with cutmix augmentation and CBAM attention for compound emotion recognition

Nidhi[1] · **Bindu Verma**[1]

Accepted: 6 June 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
Facial emotion recognition has become a complicated task due to individual variations in facial characteristics, as well as racial and cultural variances. Different psychological studies show that there are complex expressions other than basic emotions which are made up of two basic emotions like"Happily Disgusted", "Happily Surprised", "Sadly Surprised", etc. Compound emotion recognition is challenging due to very less publicly available compound emotion datasets which are imbalanced too. In this paper, we have proposed an LSwin-CBAM for the classification of compound emotions. To address the problem of the imbalanced dataset, the proposed model exploits the cutmix augmentation technique for data augmentation. It also incorporates the CBAM attention mechanism to emphasize the relevant features in an image and swin transformer with fewer swin transformer blocks which leads to less computational complexity in terms of trainable parameters and improves the overall classification accuracy as well. The experimental results of LSwin-CBAM on RAF-DB and EmotioNet datasets show that the proposed transformer-based network can well recognize compound emotions.

Keywords Swin transformer · CutMix augmentation · CBAM · Compound emotion recognition

1 Introduction

Face expressions are an efficient way for people to communicate their emotions. It gives the spectator a plethora of social cues, such as the viewer's focus of attention, emotion, motivation, and intention and brings out 55% of the conveyed message, which is relatively higher than the part communicated by the combination of voice and language [1]. Research done in social psychology explains that facial expression helps in coordinating the conversation between speaker and listener that can perceive the interest of the listener [2, 3]. In most of the studies, macro-expressions (happiness, surprise, sadness, anger, fear, and disgust) are subjects of concern that are most commonly observed in most cultures. But, sometimes humans can express mixed emotions. For example, a person can have feelings of "sad" and "surprise" components together which forms a compound emotion named "Sadly Surprised". The combination of two or more emotions is represented by a compound emotion where one emotion is dominant and the other is complementary. Face recognition in the real environment is a challenging task due to the variation in head orientation, illumination conditions, and head poses. For compound emotion recognition where dominant and complementary emotions need to be recognized, the task becomes more difficult [4].

Authors usually focus on seven basic emotions: 'Surprise', 'Anger', 'Happiness', 'Contempt', 'Fear', and 'Disgust'. But there is a need to examine more detailed and precise facial expressions. Some authors tried to find out detailed facial expressions because of recent developments in the area of compound emotions [5, 6]. To examine a person's emotional state in greater detail using facial emotion expression analysis, compound emotion categories have been proposed [7]. Such studies contribute in the context of compound emotion recognition which comprehends and identifies fine-grained facial expressions.

Transformers [8] are greatly used in Natural Language Processing (NLP) tasks. Research efforts to adapt transformers for vision tasks have been driven by the tremendous achievement of transformers in NLP. Due to its original transformer architecture and promising performance in vision

✉ Bindu Verma
bindu.cvision@gmail.com

Nidhi
nidhi.kanwar189@gmail.com

[1] Department of Information Technology, Delhi Technological University, New Delhi 110042, New Delhi, India

Published online: 14 June 2024

tasks, Vision Transformers (ViTs) [8] received a lot of attention from AI researchers. Apart from its promising performance, there are a few limitations encountered in ViTs. The computational complexity of ViTs is quadratic to image size, which is why it particularly faces difficulties with high-quality input images. Moreover, the variable scale visual features in vision tasks make it unsuitable as it processes the fixed scale tokens.

Liu et al. [9] developed the swin transformer which is possibly the most intriguing concept of research following the original ViT. Swin transformer addressed the limitations of ViT by introducing these key concepts: hierarchical feature maps and shifted windows. The key idea behind the swin Transformer is to divide the input image into non-overlapping patches and organize them hierarchically. Unlike traditional vision transformers that operate on fixed-size patches, the swin Transformer employs shifted windows, which means that each patch is constructed by shifting a fixed-size window across the image. This allows the model to have a larger receptive field without increasing the computational complexity significantly. It outperforms ViTs in image classification by identifying fine-grain features in an image where each pixel is labeled as in semantic segmentation. The deep swin transformer model is modified by including a single stage only with additional modules which improves the classification accuracy with minimal computational overhead. The adoption of lightweight machine learning models offers several significant benefits for real-time applications. Lightweight models are designed to have fewer parameters and require less computational power compared to heavy models with more computation. Due to their reduced complexity, lightweight models often have faster inference times, allowing for real-time processing of facial expressions. Training lightweight models typically requires less time and resources compared to training larger models. This can accelerate the development and deployment process, allowing for quicker adaptation to changing requirements or datasets.

Next, the absence of significant, publicly available labeled datasets for compound emotions is one of the major barriers to advancing research on the automatic recognition of compound emotions. To overcome the above-mentioned problem, the CutMix data augmentation technique has been used in the model which increases the training set and solves the problem of information loss and inefficient behavior of other regional dropout methods. The Cutmix technique picks random patches from the training images and ground truth labels are combined in the ratio corresponding to the area of patches in the images. The neural network's ability to focus on specific regions, eliminate irrelevant information, and effectively extract essential features from images is made possible by using attention mechanisms. In the proposed model, the convolutional block attention module (CBAM) attention mechanism [10] has been used which sequentially combines the channel attention module to exploit the inter-channel relationship by giving varying significance to the channels of feature maps and spatial attention module to determine which parts of a feature map are more imperative. The key points of the proposed framework are as follows:

- *T*o design a lightweight compound emotion recognition model named LSwin-CBAM that is based on a streamlined swin transformer exploiting CutMix augmentation technique and CBAM attention mechanism to improve the classification rate of compound emotions.
- *C*BAM attention mechanism is incorporated to enable the network to focus on relevant regions resulting in pertinent feature extraction and introduces a trade-off between model accuracy and parameter count in the proposed model.
- *A*dditionally, an analysis of the computational efficiency and model complexity of LSwin-CBAM is conducted, providing insights into its practical applicability.

The organization of the manuscript is as follows: Section 2 reviews the state-of-the-art methods in the field of compound emotion recognition. Section 3 includes the architectural and formulation overview of the proposed method. The experimental setup and results of the proposed method are depicted in Section 4, At last, the manuscript is concluded in Section 5.

2 Literature survey

Humans can easily recognize emotions accurately, but making a fully automated machine to recognize facial expressions is still challenging. If this capability of emotion recognition can be empowered in robots or computers then robotic applications can be developed to understand the emotions of a human with a limited understanding of the relationship between emotional expressions and body movements [11]. The facial expression recognition method is developed by deforming a few control points of the FACS-based facial motion proposed by Ekman [12]. Many studies exist [13–16] which focused on seven basic expressions but to enrich more details of human emotions, compound emotions came into the picture. Jarraya et al. [17] developed a system for compound emotion recognition for autistic children in a time of meltdown crisis. Researchers assessed deep spatio-temporal geometric features of involuntary facial expressions in the time of meltdown crisis. They analyzed the relevance of different feature selection methods like filter methods, Wrapper methods, and (Neighborhood component analysis) NCA method and observed to achieve the best accuracy i.e. 85.5% using RNN classifier and Information Gain feature

selection method [17]. Guo et al. [4] compared the top winner's methods (used CNN) from FG 2017 workshop on an iCV-MEFED dataset for the recognition of compound emotion and observed that Ist model given by [18] has performed better with less misclassification rate (0.793 for validation set and 0.802 for test set). Li et al. [19] proposed a deep locality-preserving convolutional neural network (DLP-CNN) approach which seeks to maximize inter-class dispersion while maintaining local relationships to improve the discriminative ability of deep features. The performance of DLP-CNN was evaluated on the RAF-DB dataset [19] (includes basic and compound emotions) and other different benchmarks datasets.

Kaminska et al. [20] proposed a two-stage recognition model for compound emotion. The first stage introduces coarse recognition that merges appearance-based features extracted using DCNN and facial-point features and the second stage introduces fine recognition using a binary classifier. Authors [20] further tried the ensemble model which includes a one-stage and two-stage model and eventually improves the recognition rate from 19.28% to 21.83%. Shaila et al. [21] used the MobileNet model with depthwise convolution operation for compound emotion recognition. Liang et al. [22] proposed a fine-grain dataset named "FG-Emotions" to draw a more detailed analysis of facial emotions. FG-Emotions dataset has 33 fine-grained emotion classes. Authors [22] developed a "two-stream multi-scale AU-based Network (MSAU-Net)" for the prediction of fine-grained emotions. Pons et al. [23] proposed a Selective Joint Multitask approach (SJMT) which considers selective loss function (sigmoid cross-entropy) for each dataset. This approach [23] is beneficial while dealing with different datasets containing unlabeled images. The proposed loss function is incorporated with ResNet-50 and VGG16 model and found more effective when applied to datasets collected in an unconstrained environment (Oulu-Casis, SFEW) and compound emotion recognition (EmotioNet) as compared with the individual ones. Kirillov et al. [24] proposed a foundation model SAM (Segment Anything Model) using prompt engineering which solves downstream segmentation challenges on unseen data during training.

Dosovitskiy et al. [25] introduced a vision transformer that allows models to understand image structure independently as input images and are depicted as sequences and predict the class labels for the given image. The input images are processed as a sequence of patches, where each patch has been flattened into a single vector, further concatenated by the channels of all of its pixels before projecting them linearly to the defined dimension of input. Mu et al. [26] proposed a model based on visual transformers with feature fusion which is tested on in-the-wild expression datasets like RAF-DB, AffectNet, and FERPlus. Lian et al. [27] assessed GPT-4 with vision for Generalized Emotion Recognition (GER) tasks including tweet sentiment analysis, facial emotion recognition, visual sentiment analysis, micro-expression recognition, multimodal emotion recognition, and dynamic emotion recognition. Zhu et al. [28] proposed a unified framework based on a multi-branch vision transformer for multi-task facial emotion recognition and mask-wearing classification. Dong et al. [29] introduced a novel loss function called bi-center loss, which is an extension of center loss. This new function is designed to capture compound emotion features by utilizing the centers of basic emotions. Xue et al. [30] used the vision transformer model to produce relation-aware local representation and the relation between them whereas Multi-head Attention Dropping (MAD) is also incorporated in the proposed model to drop the attention maps. Multi-head self-attention joins the relevant features using information subspaces at various locations. But different self-attention may produce identical relations which can be alleviated using Multi-head Self-attention Dropping (MSAD). Vision transformer performed well for image classification tasks but has some limitations such as higher computational complexity as it computes self-attention operation globally, and creates feature maps of a single low-resolution image. Liu et al. [9] developed a Swin transformer that merges image patches in deeper layers to produce hierarchical feature maps and it computes self-attention operation within the local window which leads to linear computational complexity. Zhao et al. [31] proposed a swin transformer-based model that accepts multimodality inputs to recognize the micro-expressions. Xue et al. [32] proposed a Coarse-to-fine Cascaded Network (CFC) with Smooth Predicting (SP) and improved the classification rate by extracting both unique and universal features and by ensembling swin transformer with ResNet for the training of the model. The proposed LSwin-CBAM framework is based on a lightweight swin transformer which consists of only two consecutive swin transformer blocks with CBAM attention mechanism which improves the recognition capability of the network with very less computational overhead.

3 Proposed work

Compound emotions are the composite of basic emotions that gives a detailed understanding of a person's feeling. Transformers performed tremendously well for natural language processing (NLP) tasks. So researchers [25] developed a variant named "Vision Transformer" based on transformer architecture which can be applied for vision tasks. However, there are some limitations as it works on a constant scale. In contrast, the scale of visual entities in the given images can have large variations and computational complexity is quadratically increasing with image size. So, these above-mentioned limitations forced researchers [9] to develop the

Swin transformer model to extract dense information from the given images. Swin transformer works on each pixel and forms a hierarchical feature map that initiates with small-sized patches and progressively merges with neighboring patches as moves down in deeper transformer layers.

In our work, we proposed a transformer-based model for compound emotion recognition which used a streamlined swin transformer with a CBAM attention module. Swin transformer is taken as the base model with some architectural modifications. The general architecture of the Swin transformer (ST) [9] has 4 stages where the ST block is preceded by a linear embedding layer in the first stage and a patch merging layer in the last three stages which leads to a heavy deep model that increases computational overhead with 6M parameters on experimentation. To optimize the base model, a reduced number of stages (single stage) are included, resulting in minimal computational overhead with 0.15M parameters. The adoption of a lightweight model is suitable for real-time applications. Moreover, the CBAM (Convolutional Block Attention Module) attention mechanism is integrated, introducing a trade-off between model accuracy and parameter count in the proposed model. Additionally, the utilization of CutMix augmentation enhances the classification performance, surpassing that of the base Swin model while utilizing fewer parameters (0.15M). By carefully integrating these components, the LSwin-CBAM model achieves enhanced performance in recognizing compound emotions, surpassing the capabilities of individual methods.

The proposed architecture of LSwin-CBAM is shown in Fig. 1 that accepts the images as input which are forwarded to the cutmix augmentation module because imbalance data is a very common problem in facial expressions datasets which affects the performance of the model. The augmented dataset is fed to the Conv2D layer to produce a feature map which is given as an intermediate map to the CBAM attention module to enhance the feature representation of relevant features. CBAM is followed by a patch partitioning module to produce patches that are flattened to form a feature vector which will go through a linear embedding layer to convert it to an arbitrary dimension C. This intermediate feature map is further fed to the streamlined swin transformer block (consisting of a single stage only) and followed by a patch merging operation to merge two adjacent patches for downsampling of the feature map. After patch merging, global average pooling and dense layer are applied to produce the classified emotion as an output. Below we have discussed each steps of proposed work.

3.1 Input data and augmentation

Input to the proposed model is a compound emotion RGB images. Cutmix augmentation is applied to increase the training samples. Data augmentation is considered as one of the important steps of data pre-processing as it greatly enhances the variety of data accessible for training our models, all without the need to gather additional data samples. Basic techniques such as flipping, random cropping, and random rotation are frequently employed to train extensive models, particularly effective for smaller datasets and simpler problems. However, significant shifts and corruptions in data can occur when faced with the complexities of real-world scenarios. So, there are advanced augmentation techniques available like cutout [33], mixup [34], and cutmix [35]. In Cutout augmentation [33], a square-shaped region is eliminated from the image and substituted with either black or grey pixels, or filled with Gaussian noise. The eliminated regions

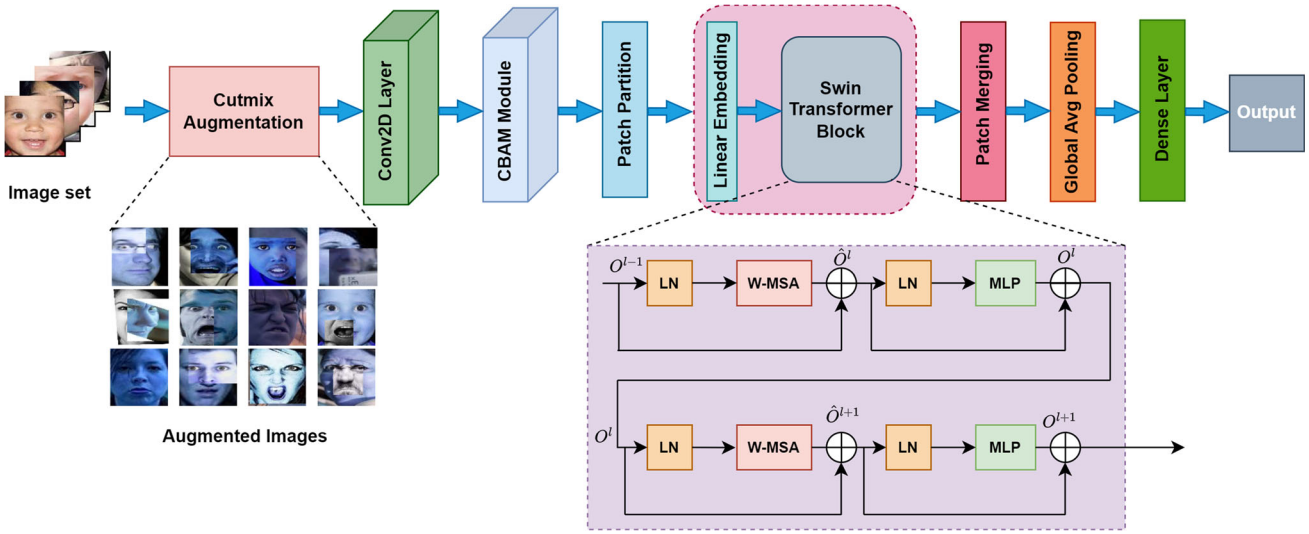


Fig. 1 Proposed model architecture integrating cutmix augmentation, CBAM attention, and lightweight Swin transformer module for compound emotion recognition

are typically zeroed out or populated by random noise [33] which drastically lowers the extent of informative pixels in given images. With mixup augmentation [34], a linear interpolation is employed between two randomly selected images. Nevertheless, the resulting images exhibit a degree of unnaturalness that can confuse the model. So, to utilize the deleted regions, Yun et al. [35] proposed the cutmix augmentation strategy. It uses a patch from a different image to restore the missing regions rather than just eliminating pixels as shown in Fig. 2. Additionally, the ground truth labels are combined in accordance with the total number of pixels in the merged images. The additional patches make the model recognize the object from a partial context, which improves the localization capability even further.

Cutmix augmentation technique [35] is applied on the given training images which produce the output images shown in Fig. 2. The training image is indicated by $x \in I^{W \times H \times C}$ where H, W, and C represent the height, width, and channels of an image respectively. Cutmix aims to produce a new training instance $(\tilde{x}, \tilde{y})$ by mixing any two given training samples (x_i, y_i) and (x_j, y_j). The model is trained using the original loss function of generated training sample $(\tilde{x}, \tilde{y})$. The combining or mixing operations are defined in (1) where $B_m \in (0, 1)^{W \times H}$ represents binary mask which suggests the pixels to keep and drop from two images.

$$\tilde{x} = B_m \odot x_i + (1 - B_m) \odot x_j,$$
$$\tilde{y} = \delta y_i + (1 - \delta) y_j, \tag{1}$$

The cutmix method will produce a cutmix-ed augmented sample $(\tilde{x}, \tilde{y})$ by combining any two randomly picked training samples in each iteration as shown in Fig. 2.

3.2 CBAM attention mechanism

The origination of simulating human attention in computer vision is to improve the model's performance for image processing tasks with minimized computational complexity. Attention mechanisms concentrate on particular sections of the images instead of the complete picture and improve the interest representation, which has gained popularity in deep learning. The convolution function extracts relevant features by incorporating spatial and cross-channel information together. The convolutional Block Attention Module (CBAM) module sequentially infers attention maps along

Fig. 2 Samples of images post-CutMix augmentation

the two distinct dimensions of channel and spatial from an intermediate feature map and then multiplies the attention maps by the input feature map to refine the adaptive feature. As a universal and lightweight module, it can be seamlessly integrated with any CNN-based architecture with minimal overhead (Fig. 3).

CBAM accepts an intermediate feature map represented by $M \in I^{C\times H\times W}$ and applies a single-dimensional channel attention map $A_c \in I^{C\times 1\times 1}$ and two-dimensional spatial attention map $A_s \in I^{1\times H\times W}$ in a sequential manner. The attention operation can be formulated in (2). In element-wise multiplication, the channel and spatial attention values are copied along spatial and channel dimensions respectively, and give M''as the final output. The visualization of learned attention weights using CBAM attention mechanism is shown in Fig. 4 to show how CBAM highlights the relevant facial features in an original image. Brighter regions indicate greater attention weights, signifying the relevance of those regions for subsequent processing.

$$\begin{aligned} M' &= A_c(M) \odot M, \\ M'' &= A_s(M') \odot M', \end{aligned} \tag{2}$$

3.3 Swim Transformer

Swin transformer [9] is developed as a general-purpose transformer-based architecture for vision tasks. It performs local self-attention within distinct windows and attains linear computational complexity. The streamlined architecture of the swin transformer used in the proposed model is illustrated in Fig. 5. Firstly, attentive features extracted from CBAM attention undergo patch partitioning process which splits input images into non-overlapping patches with a patch size of 2×2. Each patch is considered a "token," where its feature is configured as a concatenation of the pixel values of an RGB image. Each feature with dimension $2 \times 2 \times 3$ will be flattened and converted into a vector of length 12. The vector is projected to an arbitrary dimension (C) using a linear embedding layer. On these patch tokens, two swin transformer blocks (within a single stage only) with modified self-attention mechanisms are used, and maintain the number of tokens $(H/2 \times W/2)$. A transformer block with a linear embedding layer is referred to as stage 1 in Fig. 5.

3.3.1 Swin transformer block

The Swin Transformer block has a hierarchical structure, which allows it to process images of varying sizes. So, to create the Swin Transformer, the conventional multi-head self-attention (MSA) module in a transformer block is replaced with an SW-MSA module (Mutli-head attention with shifted windows) to compute the self-attention within local windows, while other layers remain the same. It includes shifted-window-based multi-head self-attention

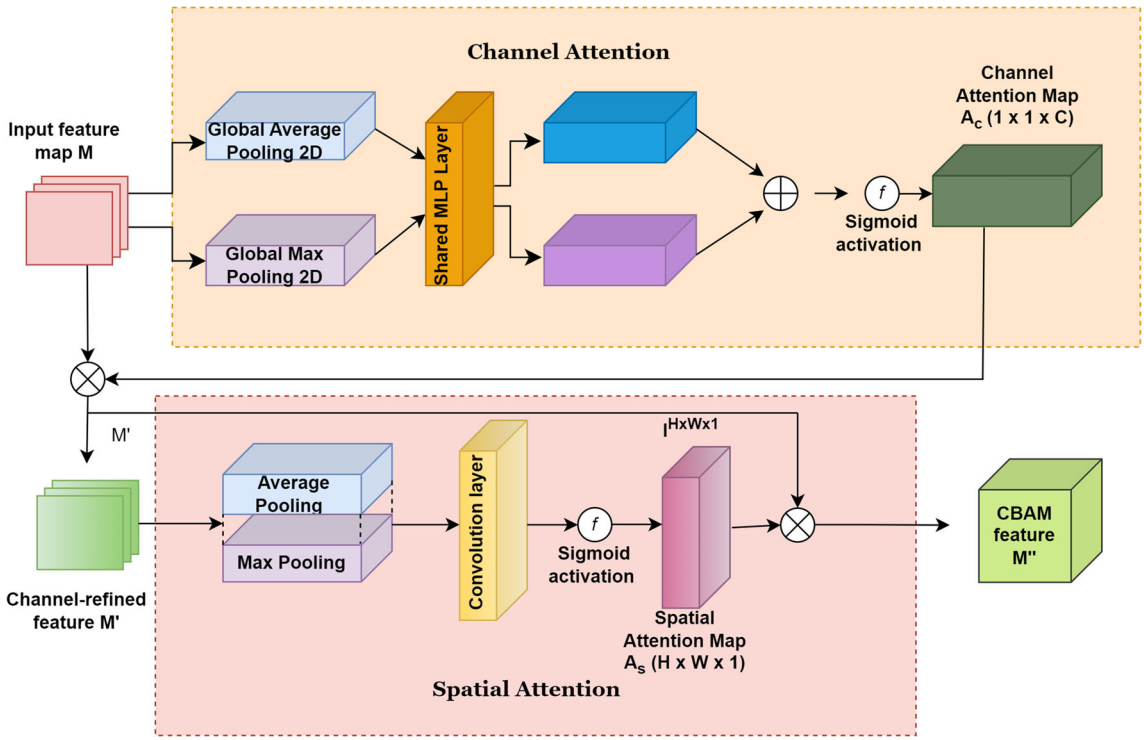


Fig. 3 Architecture of Convolutional Block Attention Module (CBAM) [10]

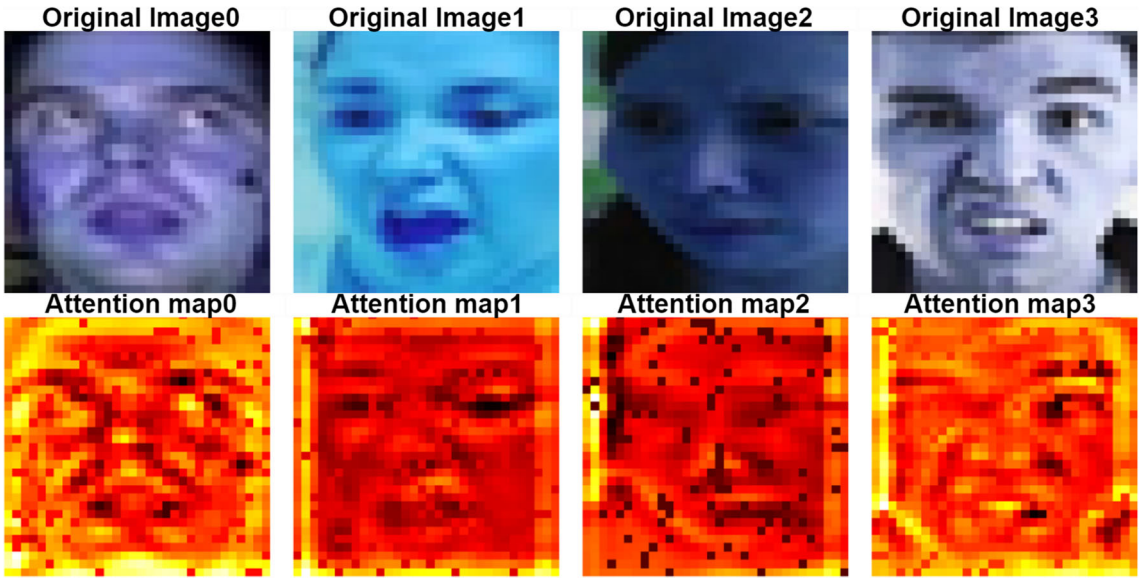


Fig. 4 Visualization of learned attention weights of CBAM attention on input images

and an MLP(2-layer) with a GELU activation function. A normalization layer is added before every MSA and MLP component, and these components are followed by the residual connection. The output of the Swin Transformer block is then passed to the patch merging layer in the network. By stacking two swin transformer blocks on top of each other, the model can learn increasingly complex representations of the input image.

In SW-MSA, the windows have been organized for image partitioning so that the partitions don't overlap with each other. This mechanism improves the computation complexity as shown in (3). Suppose the window has $W \times W$ patches, the complexity of window-based self-attention and the basic MSA module on a given image of $h \times w$ patches can be defined as:

$$\begin{aligned} \Omega(MSA) &= 4hwC^2 + 2(hw)^2C, \\ \Omega(W-MSA) &= 4hwC^2 + 2W^2hwC, \end{aligned} \tag{3}$$

A shifted Window partitioning approach is followed between successive swin transformer blocks to incorporate cross-window links and maintain the computation efficiency of non-overlapping windows at the same time. With this approach, the output of successive swin transformer blocks is defined as:

$$\begin{aligned} \hat{O} &= W-MSA(LN(O^{l-1})) + O^{l-1}, \\ O^l &= MLP(LN(\hat{O}^l)) + \hat{O}^l, \\ \hat{O}^{l+1} &= SW-MSA(LN(O^l)) + O^l, \\ O^{l+1} &= MLP(LN(\hat{O}^{l+1})) + \hat{O}^{l+1}, \end{aligned} \tag{4}$$

W-MSA denotes window-based self-attention and SW-MSA denotes shifted windows-based multi-head self-atten tion. In the above (4), $\hat{O}^l$ and O^l represent the output of SW-MSA and MLP modules for l block, respectively.

3.4 Compound emotion classification

The output of the swin transformer block is fed to the patch merging layer to merge two adjacent patches for downsampling of the feature map. This layer is followed by a global

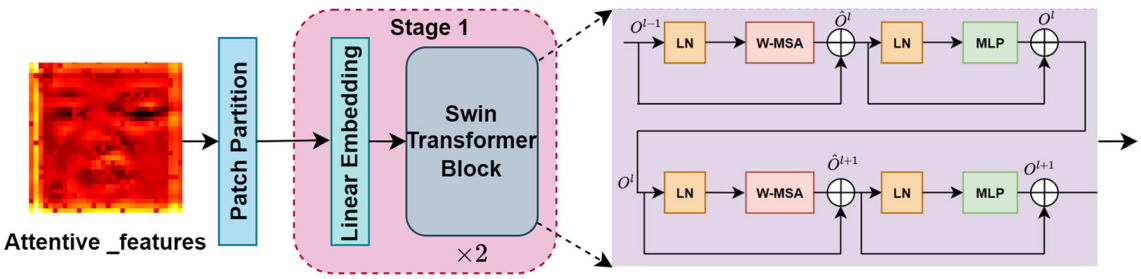


Fig. 5 Streamlined swin transformer architecture used in the proposed model

averaging layer to reduce the dimensionality of the output, making it computationally efficient while preserving relevant information for downstream tasks. The output of the global averaging layer serves as input to a dense layer. In a dense layer with softmax activation, the output is produced by applying the softmax function on outputs of the global average pooling layer s_j by applying the softmax function to obtain probabilities ($\mathbb{Y}$) for each class(i) as shown in (5). The categorical cross-entropy loss (CE) is computed using the predicted probability $\mathbb{Y}_i$ and true probability ditribution Y_i for the i^{th} class in (5).

$$\mathbb{Y}_i = \frac{e^{s_i}}{\sum_j^{\mathbb{C}} e^{s_j}}$$
$$CE = \sum_i^{\mathbb{C}} Y_i log(\mathbb{Y}_i) \tag{5}$$

The provided algorithm 1 outlines the processing steps for training a compound emotion recognition model, referred to as LSwin-CBAM, using a dataset (RAF-DB, EmotioNet). The algorithm starts by taking a dataset $\{X_i, Y_i\}$ n, where X_i represents facial images, and Y_i represents corresponding labels for compound emotions. The cutmix augmentation method is applied to the dataset, creating augmented samples $\{\hat{x}_i, \hat{y}_i\}$. The CBAM (Convolutional Block Attention Module) attention mechanism is applied to enhance the model's ability to focus on important regions within facial images. A 2D convolutional layer is used to produce a feature map M from the augmented images and further processed through channel and spatial attention operations producing M' and M'' output respectively. The feature map M'' is partitioned into patches of size P_s. The algorithm proceeds through the stage of swin transformer block which involves a linear embedding layer, iterative application of swin Transformer blocks $\times 2$ followed by a patch merging layer(generally applied, if the current stage is not the last one, but here single stage is used and patch merging layer is also applied exceptionally). Emphasizing efficiency without compromising accuracy, this model enables real-time processing, making it ideal for deployment in resource-constrained environments while maintaining a high level of precision in emotion detection.

Algorithm 1 Algorithm of the proposed model.

Input:

Dataset $\in \{X_i, Y_i\}_{i=1}^n$ where n is number of images in dataset
Model parameters θ
Initial learning rate l_r
Batch Size B
Dropout Dp
Learning rate decay factor d_r after every n epochs
Number of epochs $\mathbb{E}$
Patch size $\mathbb{P}_s$
Number of stages $\mathbb{N}_s$
Depth of Swin transformer block $\mathbb{D}$

Output:

Trained LSwin-CBAM model to recognize compound emotions from facial images

Input the dataset, where Dataset $\in \{X_i, Y_i\}_{i=1}^n$
Apply the cutmix augmentation method
for `i = 1 to n` **do**
    $\{\hat{x}_i, \hat{y}_i\} = cutmix(X_i, Y_i)$,
end for
Apply CBAM attention mechanism
    Apply a 2D convolutional layer to produce a feature map M
    Output of Channel attention: $M' = A_c(M) \odot M$,
    Output of Spatial attention: $M'' = A_s(M') \odot M'$,
Apply patch partitioning to partition the patches of size $\mathbb{P}_s$
Stages:
    Apply linear embedding layer
for `i = 1 to` $\mathbb{N}_s$ **do**
    for `j = 1 to` $\mathbb{D}$ **do**
        Apply swin transformer block
        if $\mathbb{N}_s = 1$ **or** $i < \mathbb{N}_s - 1$ **then**
            Apply patch merging layer
        end if
    end for
end for
Apply the global average pooling layer
Apply the dense layer which gives the final output

4 Experimental results

The system requirements to evaluate the proposed model are 64-bit Windows 11, Intel Core i7 processor with 16GB RAM size, 8GB NVIDIA TITAN RTX graphics card. Tensorflow 2.8 has been used for the implementation of the proposed framework. The selection of hyperparameters is indeed a crucial aspect of our model development. The summary of hyperparameters of the model is shown in Table 1. All the hyperparameters used in the model are decided experimentally. These hyperparameters collectively define the configuration of a proposed model for facial emotion recognition, and these values have been meticulously selected through multiple experiments to determine the performance and behavior of the model during training. The learning rate is set to 0.001, this parameter is decided experimentally and controls the step size during the optimization process with learning rate decay of 0.0001. The dropout rate is 0.05 to prevent the model from overfitting by dropping several neurons. Several experiments were conducted using different dropout values, and the results, presented in Tables 2 and 3, indicate that the optimal value chosen is 0.05. The best results are highlighted with bold text. AdamW, is used for training LSwin-CBAM due to its improved weight decay handling

Table 1 Summary of hyperparameters used in LSwin-CBAM

Hyperparameter	Value
Batch Size	16
Learning rate	0.001
Learning rate decay factor	0.0001
Dropout	0.05
Number of Epochs	500
Patch Size	2×2
Window Size	2×2
Number of Stages	1
Depth of Swin transformer block	2

(weight decay = 0.0001). Improved weight decay handling in AdamW contributes to reduce the generalization gap, making the model generalize better to unseen facial expressions and diverse datasets. Along with that, AdamW has been observed to provide more stable training and faster convergence compared to standard Adam, which is crucial for effectively training complex models like transformers used in facial emotion recognition. Categorical Cross-Entropy (CCE) loss is used for Transformer models in multi-class classification due to its ability to compare predicted probability distributions with true labels, providing smooth gradients for optimization.

Table 2 Experimental results of LSwin-CBAM on RAF-DB dataset, * indicates results of model trained with class weights

Batch_size	Dropout	Acc	RA	P	R	$F1_m$
16	0.04	49.76%	0.78	0.33	0.28	0.27
	0.05	**51.81%**	**0.78**	**0.45**	**0.30**	**0.30**
	0.06	50.12%	0.78	0.34	0.28	0.26
	0.07	49.64%	0.79	0.40	0.29	0.29
32	0.04	49.17%	0.77	0.32	0.30	0.29
	0.05	50%	0.77	0.35	0.29	0.27
	0.06	50.36%	0.77	0.36	0.29	0.29
	0.07	47.51%	0.76	0.25	0.25	0.23
64	0.04	47.98%	0.76	0.44	0.26	0.25
	0.05	48.46%	0.76	0.35	0.29	0.27
	0.06	46.56%	0.77	0.38	0.25	0.23
	0.07	48.34%	0.75	0.39	0.28	0.28
128	0.04	47.03%	0.76	0.37	0.26	0.26
	0.05	47.62%	0.76	0.38	0.29	0.29
	0.06	45.96%	0.74	0.40	0.26	0.25
	0.07	46.67%	0.77	0.37	0.29	0.29
*16	0.05	48%	0.74	0.36	0.31	0.31

Table 3 Experimental results of LSwin-CBAM on EmotioNet dataset, * indicates results of model trained with class weights

Batch_size	Dropout	Acc	RA	P	R	$F1_m$
16	0.04	36.38%	0.56	0.13	0.10	0.09
	0.05	39.67%	0.59	0.21	0.09	0.09
	0.06	35.76%	0.59	0.13	0.09	0.09
	0.07	36.38%	0.60	0.07	0.08	0.07
32	0.04	35.56%	0.59	0.12	0.09	0.09
	0.05	36.59%	0.58	0.15	0.11	0.11
	0.06	35.55%	0.57	0.15	0.09	0.10
	0.07	37.21%	0.61	0.24	0.12	0.11
64	0.04	37.42%	0.61	0.18	0.13	0.12
	0.05	36.17%	0.61	0.19	0.11	0.12
	0.06	36.38%	0.61	0.11	0.10	0.10
	0.07	36.38%	0.61	0.13	0.11	0.10
128	0.04	35.97%	0.58	0.10	0.09	0.07
	0.05	35.34%	0.60	0.18	0.13	0.13
	0.06	34.3%	0.59	0.10	0.10	0.14
	0.07	34.72%	0.59	0.18	0.11	0.11
*16	0.05	32.26	0.56	0.09	0.08	0.06

4.1 Datasets

4.1.1 Realworld Affective Face Database (RAF-DB)

The RAF-DB (Real-world Affective Faces Database) [19] is a face expression dataset. It contains 29672 facial photos labeled by 40 different taggers with seven basic ('Anger', 'Sadness', 'Happiness', 'Disgust', 'Fear', 'Surprise', and 'Neutral') and eleven compound expressions (0: 'Happily Surprised', 1: 'Happily Disgusted', 2: 'Sadly Fearful', 3: 'Sadly Angry', 4: 'Sadly Surprised', 5: 'Sadly Disgusted', 6: 'Fearfully Angry', 7: 'Fearfully Surprised', 8: 'Angrily Surprised', 9: 'Angrily Disgusted', 10: 'Disgustedly Surprised'). The participants' age, gender, and ethnicity, as well as lighting conditions, head postures, occlusions (e.g. spectacles, self-occlusion, or facial hair), and post-processing processes (e.g. special effects and various filters), are all represented in this database. The number of samples in each class is depicted in Fig. 6 which shows that the dataset is highly imbalanced.

4.1.2 EmotioNet

EmotioNet [7] is a large facial expression dataset that has nearly one million facial images and corresponding labels. It offers a wide range of facial expressions including basic and compound expressions. Images are downloaded from

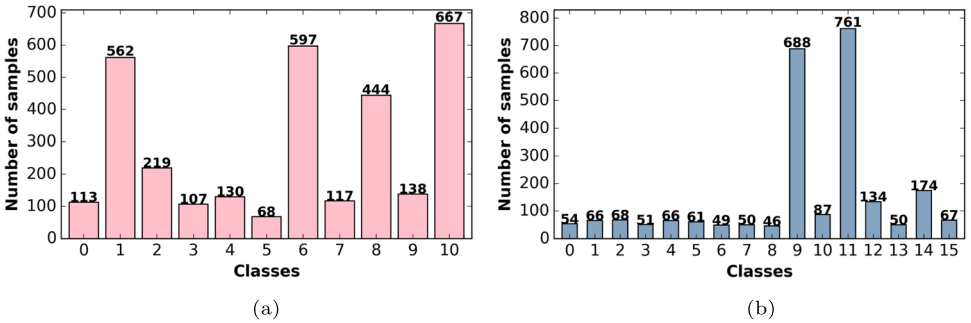


Fig. 6 Number of samples in a) RAF-DB dataset b) emotionet dataset

the internet by using different keys for 23 emotions based on action units (AUs). In our experiment, 2474 images are used which has 16 basic and compound emotion categories (0: 'Angrily disgusted', 1: 'Angrily surprised', 2:'Angry', 3: 'Appalled', 4: 'Awed', 5: 'Disgusted', 6: 'Fearful', 7: 'Fearfully angry', 8: 'Fearfully surprised', 9: 'Happily disgusted', 10: 'Happily surprised', 11: 'Happy', 12: 'Sad', 13: 'Sadly angry', 14: 'Sadly disgusted', 15: 'Surprised'). The EmotioNet dataset is highly imbalanced due to the presence of skewed class proportion as shown in Fig. 6

The proposed model is evaluated on two compound emotion datasets: RAF-DB and EmotioNet. Both datasets are imbalanced datasets as their class distribution is very unequal which significantly affects the performance of the model.

4.2 Class weights

Class imbalance is a common problem in image/video classification problems. It merely indicates that the frequency of a class is severely unbalanced, i.e., the frequency of one class is far more than that of the other classes. In other words, it introduces a bias towards the majority classes. To overcome this problem, class weights are adopted which assign weights to the classes by giving less weight to majority classes and high weights to minority classes which penalized the misclassification rate of the model. The general formula to calculate the class weight for each given class is shown in (6).

$$W_i = N_i/(C * N_i), \tag{6}$$

where W_i denotes weight for i^{th} class, N_i and C represents number of samples in i^{th} class and number of classes respectively.

4.3 Evaluation metrics

As RAF-DB and EmotioNet datasets are imbalanced, accuracy is not considered a good metric for evaluating the model. This can be understood by a simple example. Suppose a dataset has two classes: Real or Fake and Real class has 1000 samples (majority class) whereas the Fake class has only 10 samples (minority class). After training on the given dataset,

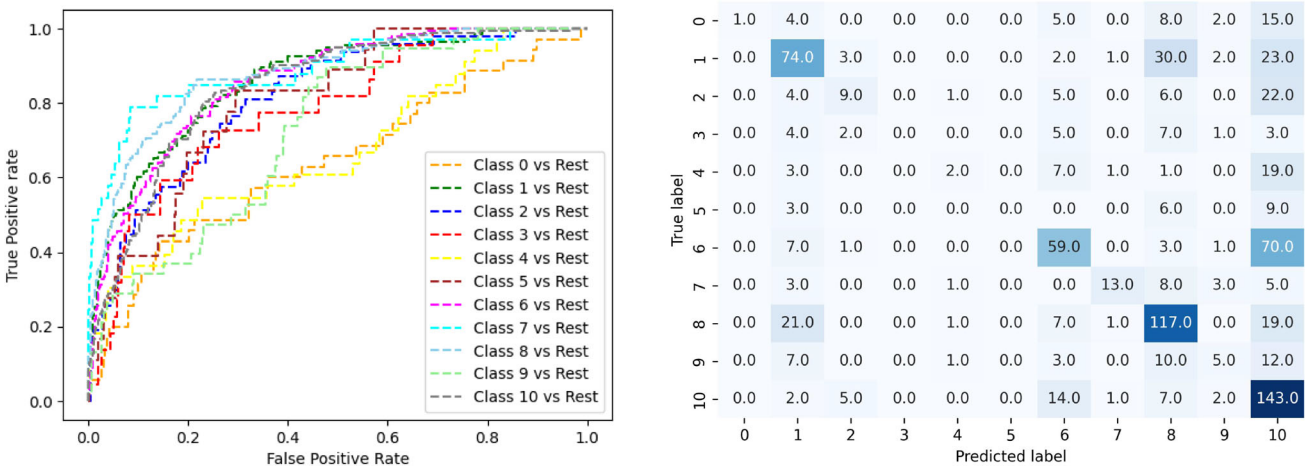


Fig. 7 Evaluated results on RAF-DB dataset a) ROC_Curve b) Confusion metric

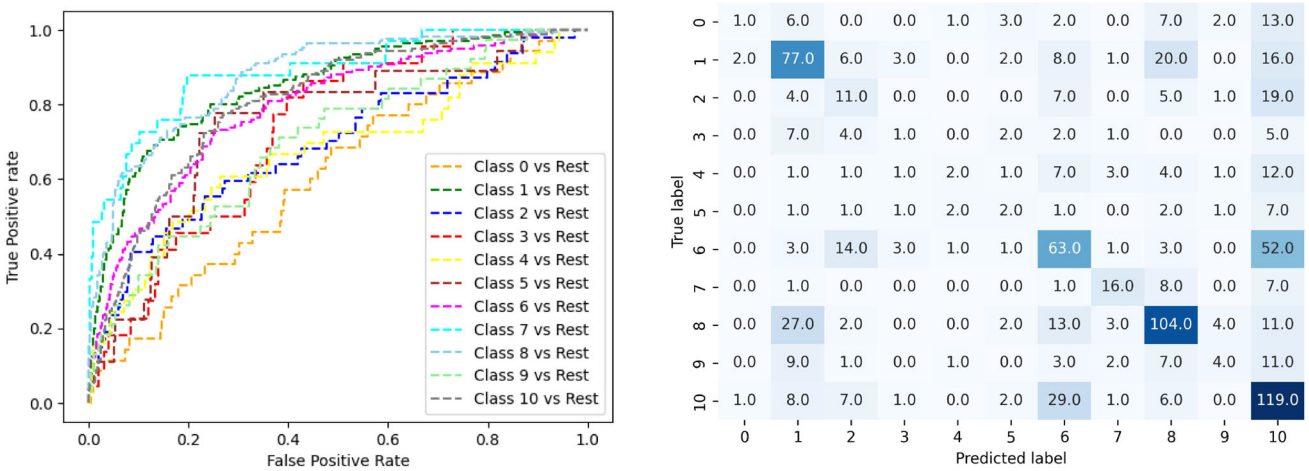


Fig. 8 Evaluated results on RAF-DB dataset (with class weights) a) ROC_Curve b) Confusion metric

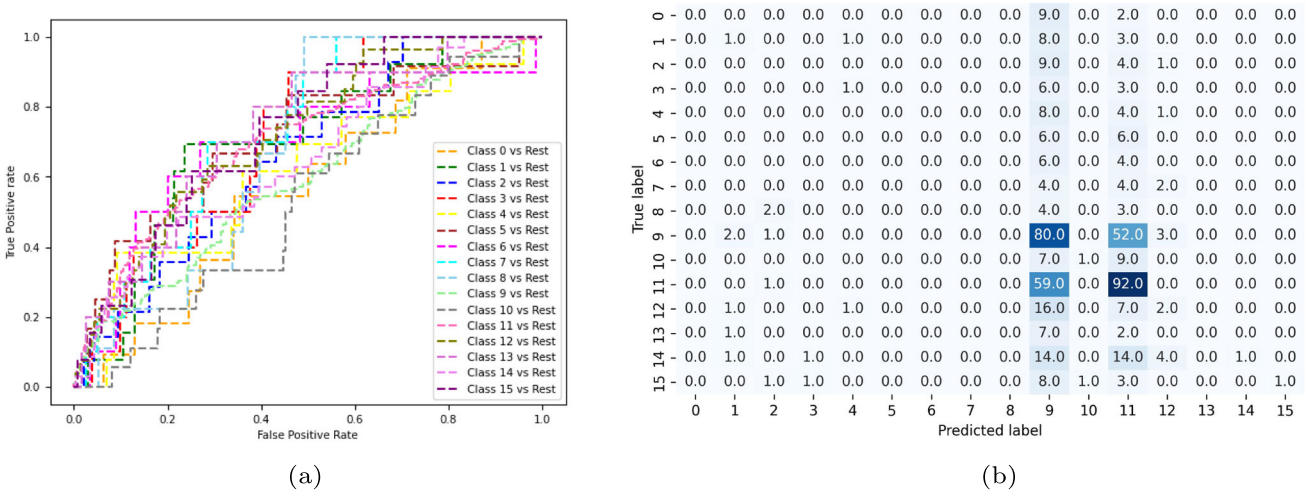


Fig. 9 Evaluated results on Emotionet dataset a) ROC_Curve b) Confusion metric

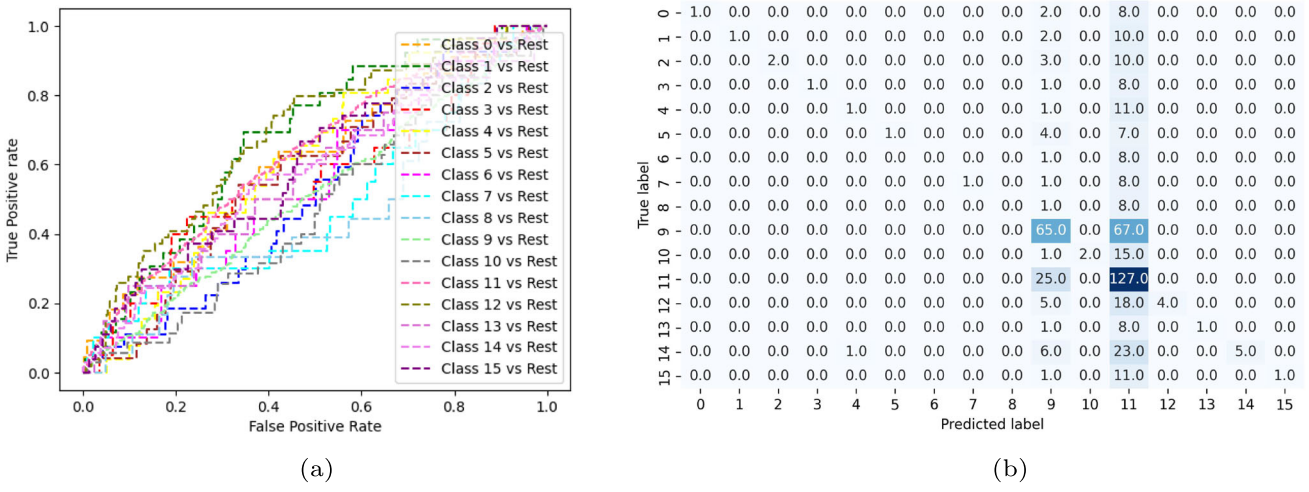


Fig. 10 Evaluated results on Emotionet dataset (with class weights) a) ROC_Curve b) Confusion metric

Table 4 Significance of adding each module to LSwin-CBAM while experimenting on RAF-DB dataset

Model	Acc	AUC	P	R	$F1_m$
Baseline Swin	45.25%	0.73	0.41	0.27	0.26
LSwin+Cutmix	47.22%	0.77	0.41	0.31	0.33
LSwin-CBAM	**51.81%**	**0.78**	**0.45**	**0.30**	**0.30**

the model gives 99% accuracy (predicting 1000 samples correctly including 999 real and 1 fake sample or Real: 99% accuracy and Fake: 1%) which is remarkable indeed. If we turn the spotlight on these results, it signifies that the model failed to predict the fake samples correctly which concludes that the proposed model is not a good predictor. So, it would be unfair to consider the accuracy as a performance metric for imbalanced datasets. Hence, other evaluation metrics are considered to justify the proposed model like precision (P), recall(R), f1 score (F1), ROC curve (Receiver operating characteristic curve), ROC AUC score (Area under the Receiver Operating Characteristics Curve). ROC curve and ROC AUC Score (AUC) have been considered as the optimistic metric to evaluate the performance of the classification model on imbalanced datasets because it is independent of class distribution and invariant to the threshold.

4.4 Experimental results on RAF-DB(Compound) dataset

The RAF-DB dataset contains 3240 images in the training subset and 840 images in the testing subset [19]. The proposed model is then trained with different configurations changing batch size, dropout, and learning rate whose results are shown in Table 2. The cosine learning rate decay is adopted to avoid the problem of local minima. Firstly, we experimented with the model by taking different batch sizes and dropout giving the best results with 16 and 0.05 in terms of accuracy(Acc), ROC_score, precision, recall, and f1-score (51.81%, 0.78, 0.45, 0.30, and 0.30 respectively). Through different experiments, it has been observed that increasing batchsize degrades the performance of the model. We also examined the influence of dropout on the model performance on different values of batchsize constant (16, 32, 64, 128) and found the optimal value of dropout which is 0.05. The value

Table 5 Significance of adding each module to LSwin-CBAM while experimenting on EmotioNet dataset

Model	Acc	AUC	P	R	$F1_m$
Swin	32.12%	0.55	0.13	0.10	0.11
LSwin+Cutmix	33.95%	0.57	0.10	0.07	0.6
LSwin-CBAM	**39.67%**	**0.59**	**0.21**	**0.09**	**0.09**

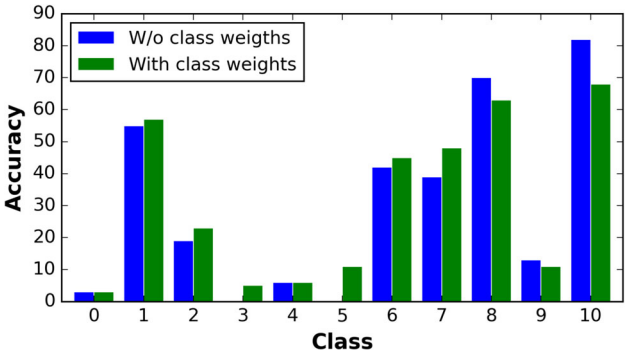


Fig. 11 Class-wise average accuracy obtained on RAF-DB without and with class weights

of dropout above 0.05 shows degraded results on the dataset as shown in Table 2.

The model gives 51.81% accuracy but due to an imbalanced dataset, it cannot classify minority classes (having a very less number of samples in the training and testing subset). To overcome this problem, class weights are introduced during the training process so that the model emphasizes the minority class to improve the classification rate of minority classes as well. The class weights are initialized according to the formula given in (6) but with multiple experiments using different weight sets, the optimized weight set is [0: 2.60, 1: 0.92, 2: 1.744, 3: 2.95, 4: 2.265, 5: 4.331, 6: 0.83, 7: 2.517, 8: 0.963, 9: 2.134, 10: 0.841]. The model with class weights classified the testing samples with an accuracy of 48% which is lowered but class weights are exploited to make the model learn all the classes (including the minority classes) and classify them as shown in Figs. 7 and 8. Figure 8a) shows that class weights improved the AUC of minority classes which is lower in Fig. 7a) due to imbalanced datasets. Figure 8b) shows how class weights improved the corrected classified samples of minority classes as well which are lesser or zero in the confusion metric shown in Fig. 7b).

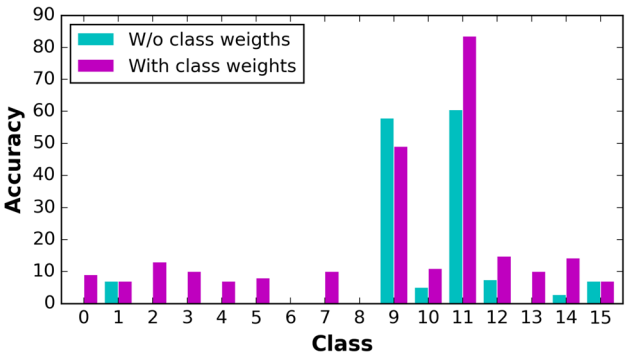


Fig. 12 Class-wise average accuracy obtained on Emotionet without and with class weights

Table 6 Parametric and architectural influence of swin transformer stages on the performance of model on RAF-DB dataset

Model	Stages	Param	Acc	AUC	P	R	$F1_m$
Swin4-CBAM	4	6M	49.29%	0.78	0.43	0.29	0.28
Swin3-CBAM	3	2.9M	49.41%	0.78	0.39	0.28	0.28
Swin2-CBAM	2	0.41M	49.64%	0.77	0.35	0.27	0.26
LSwin-CBAM	**1**	**0.15M**	**51.81%**	**0.78**	**0.45**	**0.30**	**0.30**

4.5 Experimental results on EmotioNet dataset

In our experiment, a set of emotionet challenges having 2474 facial expression images have been used depicting basic and compound emotions [7]. The images are manually annotated but these are a very small number of images for the training of the model. The dataset is split in the ratio of 80:20 giving 1924 images to the training set and 481 to the testing set. With extensive experiments, the proposed model has achieved 39.67% accuracy. To the best of our knowledge, there is no manuscript available that has considered a small set of the EmotioNet dataset for training purposes. The results are obtained by varying different batch sizes and dropouts as shown in Table 3. Due to an imbalanced dataset, the minority classes are not well classified so class weights are incorporated and the performance of the model without and with class weights is depicted in Figs. 9 and 10. By comparing Figs. 9a) and 10a), it is clear that adding class weights has improved the AUC of minority classes. In the confusion metric shown in Fig. 9b), minority classes are not predicted well but adding class weights makes the model recognize the minority classes as well, shown in diagonal entries of confusion metric in Fig. 10.

The imbalance in the Emotionet dataset is what prevents it from accurately predicting minority classes. There are different methods available to handle imbalanced datasets. Class weights are used here to reduce the biases of the model towards the majority class by assigning higher weights to the minority class which affects the overall performance but improves the class-wise accuracy of the minority classes.

4.6 Ablation study

The ablation study is provided in this section to justify the reasonability of the proposed model. Quantitative evaluation is conducted to support the contribution of each module in the model. Initially, a swin transformer has experimented alone on RAF-DB and Emotionet datasets which classify the facial expression with 45.25% and 32.12% accuracy respectively. As the given datasets are imbalanced, adding the cutmix augmentation technique improved the performance by 3.02% (RAF-DB), 1.75% (EmotioNet) and hence classifies the basic and compound emotions with 47.22%(RAF-DB), and 33.95%(EmotioNet) accuracy. The CBAM attention mechanism emphasizes channel and spatial information, enhances the target regions, and extracts relevant and detailed information from the given images. Adding this module improved the classification performance improved by 4.59%(RAF-DB), 5.72%(EmotioNet) which leads to 51.81% and 39.67% accuracy for RAF-DB and EmotioNet datasets respectively. Tables 4 and 5 show the results of RAF-DB and EmotioNet datasets in terms of accuracy and ROC_Score (AUC), precision (P), recall(R), and f1-score ($F1_m$) obtained by adding each module to the model. Figures 11 and 12 shows the class-wise accuracy obtained without and with class weights on RAF-DB and EmotioNet datasets respectively. Results show that incorporating class weights has improved the class-wise accuracy of both datasets.

4.6.1 Parametric and architectural influence

The general architecture of the Swin transformer (ST) [9] has 4 stages where the ST block is preceded by a linear embedding layer in the first stage and a patch merging layer in the last three stages. This complex architecture leads to a large number of trainable parameters which increases training time without inducing much improvement in classification performance. So, this section shows the ablation study of parametric and architectural influence by adding the number of stages in the model. Tables 6 and 7 shows the significance of adding each stage to the model and results show the effectiveness of the lightweight LSWin-CBAM model for compound emotion recognition.

Table 7 Parametric and architectural influence of swin transformer stages on the performance of model on Emotionet dataset

Model	Stages	Param	Acc	AUC	P	R	$F1_m$
Swin4-CBAM	4	6M	35.56%	0.57	0.08	0.08	0.06
Swin3-CBAM	3	2.9M	35.35%	0.61	0.11	0.10	0.09
Swin2-CBAM	2	0.41M	34.75%	0.58	0.10	0.09	0.09
LSwin-CBAM	**1**	**0.15M**	**39.67%**	**0.64**	**0.21**	**0.09**	**0.09**

Table 8 Comparative analysis of state-of-the-art methods and proposed method on RAF-DB Dataset

Model	Accuracy	Parameters_size(MB)
MAE [39]	24.94%	≈ 7.65
VGG + mSVM [19]	31.6%	≈ 528
CoAtNet [40]	33.42%	≈ 52.21
baseDCNN + mSVM [19]	40.2%	≈ 5
ViT [8]	44.18%	≈ 162.24
DLP-CNN [19]	44.6%	≈ 1.34
ReCNN [41]	46.1%	≈ 4.82
VGG16+PSR [42]	46.5%	≈ 528
Fine-tuned MSAU-Net [22]	50.2%	–
Zero-shot C-EXPR-NET [43]	51.7	≈ 14
LSwin-CBAM	**51.81%**	≈ 0.582

4.7 Comparative analysis of the proposed model with state-of-the arts

Firstly, the model is trained with RAF-DB and EmotioNet datasets separately. The evaluated results are compared with the state-of-the-art methods as shown in Tables 8 and 9 for RAF-DB and EmotioNet datasets respectively. Table 8 shows that LSwin-CBAM outperforms state-of-the-art methods by 1.11% the best performing Multi-Scale Action Unit (AU)-based Network (MSAU-Net), 6.71% the DLP-CNN pre-trained on RAF-DB basic emotions and fine-tuned on RAF-DB compound emotions, 5.21% the Relation Convolutional Neural Network (ReCNN) to capture the associations between relevant regions and facial expressions, 4.21% the Amending Representation Module (ARM) (a variant of Resnet-18), 3.01% the finetuned FaceBehaviorNet which is pre-trained on millions of images taken from 10 datasets for multi-task learning of basic expressions and then finetuned on RAF-DB for compound emotion recognition, 4.81% the Pyramid with super-resolution (PSR) network for in-the-wild FER. The EmotioNet dataset has 2472 images with unequal distribution of the classes (imbalanced dataset) and there is limited literature leveraging this dataset. The Table 9 illustrates the comparison of the proposed model with state-of-the-art methods for EmotioNet dataset, indicating

Table 9 Comparative analysis of state-of-the-art methods and proposed method on EmotioNet Dataset

Model	Accuracy	Parameters_size(MB)
ResNet50 [44]	35.01%	≈ 22.53
ResNet50-LSR [37]	37.23%	≈ 22.53
ResNet50-PS-KD [38]	38.95%	≈ 22.53
ResNet50-ISLM [36]	38.15%	≈ 22.53
LSwin-CBAM	**39.67%**	≈ 0.582

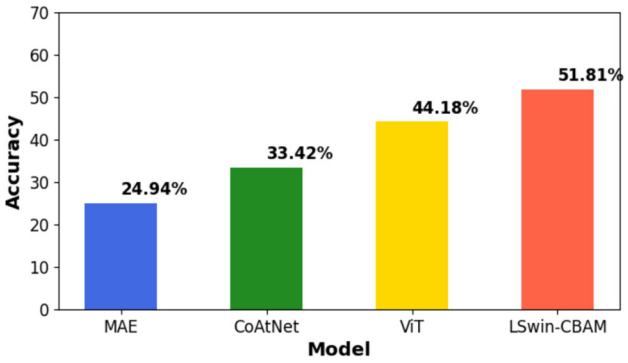


Fig. 13 Performance evaluation: LSwin-CBAM vs. Transformer Variants

its superior performance over existing models while utilizing a significantly lower number of parameters. Jiang et al. [36] used expression soft label mining (ESLM) to produce soft target labels automatically for the learning process with the ResNet50 classification model. ResNet50 was tested with LSR (Label smoothing regularization) [37] and PS-KD(Progressive self-knowledge distillation) [38], and ISLM(Iterated soft label mining [36] methods achieving classification accuracies of 37.23%, 38.95%, 38.15% respectively.In addition to the enhanced accuracy, the LSwin-CBAM model achieves a smaller parameter size of 0.581MB compared to other models (ResNet50-LSR [37], ResNet50-PS-KD [38], ResNet50-ISLM [36]) which utilize 22.53MB.

The comparison of LSwin-CBAM with transformer-based models like ViT [8], MAE [39], and COAtNet [40] highlights its superior performance in image classification of compound emotions as shown in Table 8 and Fig. 13. Its efficiency in utilizing limited training data makes it more suitable for scenarios with sparse data availability, as heavy models are prone to under-fitting. LSwin-CBAM's lightweight design also enables better scalability and adaptability to smaller datasets compared to transformer models like ViT and COAtNet, which struggle with imbalanced datasets due to their demand for extensive samples per class during training. Moreover, MAE suffers from poor reconstruction quality and instability during training, while CoAtNet and ViT struggle with insufficient representative samples for each class, resulting in suboptimal performance.

5 Conclusion and future work

This study proposed LSwin-CBAM model for compound emotion recognition that incorporates cutmix augmentation to augment the training data and CBAM attention mechanism to emphasize critical sections of the image. Experimental results on RAF-DB and EmotioNet datasets show the superiority of this model in terms of different evaluation metrics

(accuracy, auc, precision, recall, f1-score) over the state-of-the-art methods. The comparative results shown in Table 8 prove that LSwin-CBAM classified the compound emotion with good accuracy and with less computational complexity. The proposed model achieved 51.81% for RAF-DB and 39.67% for the EmotioNet dataset but fails to classify the minority classes as well. This problem occurs due to the imbalanced dataset which can be alleviated by incorporating class weights. Using class weights declined the overall accuracy but make it to classify minority classes for RAF-DB and EmotioNet datasets. The proposed model is very lightweight in terms of trainable parameters (0.15M) and still achieved impressive results which take it above the state-of-the-art methods.

This model improves the classification performance for compound emotions, but there is still scope for improvements that can be considered for future work. In future work, we would like to explore different feature extractors exclusively such as CoCa [45], OmniVec [46]. There are very less facial datasets available for compound emotions which limits the training of the network. As deep neural networks are data-hungry, the more the training data is provided, the more accurately it would classify the compound emotions. So, there is a need for balanced and large datasets for compound emotions which can improve the learning process and further improve the classification rate as well.

Acknowledgements The authors would like to acknowledge the Department of Information Technology (Cyber Forensic and Malware Analysis Lab), Delhi Technological University, New Delhi, India for providing me necessary resources to carry out the research.

Data Availability The datasets RAF-DB and EmotioNet used in this study are available to the public in original paper.

Declarations

Competing of interest On behalf of all the authors, the corresponding author declares that there is no competing financial interest to influence the work presented in this paper.

References

1. Mehrabian A (1981) Silent messages: implicit communication of emotions and attitudes. Wadsworth Pub, Co
2. Boyle EA, Anderson AH, Newlands A (1994) The effects of visibility on dialogue and performance in a cooperative problem solving task. Lang Speech 37(1):1–20
3. Stephenson G, Ayling K, Rutter D (1976) The role of visual communication in social exchange. Br J Soc Clin Psychol 15(2):113–120
4. Guo J, Lei Z, Wan J, Avots E, Hajarolasvadi N, Knyazev B, Kuharenko A, Junior JCSJ, Baro X, Demirel H et al (2018) Dominant and complementary emotion recognition from still images of faces. IEEE Access 6:26391–26403
5. Bellocchi A (2015) Methods for sociological inquiry on emotion in educational settings. Emot Rev 7(2):151–156
6. Loob C, Rasti P, Lüsi I, Jacques JC, Baró X, Escalera S, Sapinski T, Kaminska D, Anbarjafari G (2017) Dominant and complementary multi-emotional facial expression recognition using c-support vector classification. In: 2017 12th IEEE International conference on automatic face & gesture recognition (FG 2017), pp 833–838. IEEE
7. Fabian Benitez-Quiroz C, Srinivasan R, Martinez AM (2016) Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5562–5570
8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A.N, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Advances in neural information processing systems 30
9. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
10. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: Convolutional block attention module. In: Proceedings of the european conference on computer vision (ECCV), pp 3–19
11. Luo Y, Ye J, Adams RB, Li J, Newman MG, Wang JZ (2020) Arbee: Towards automated recognition of bodily expression of emotion in the wild. Int J Comput Vis 128(1):1–25
12. Ekman P (1993) Facial expression and emotion. Am Psychol 48(4):384
13. Calvo MG, Fernández-Martín A, Gutiérrez-García A, Lundqvist D (2018) Selective eye fixations on diagnostic face regions of dynamic emotional expressions: Kdef-dyn database. Sci Rep 8(1):1–10
14. Cimtay Y, Ekmekcioglu E, Caglar-Ozhan S (2020) Cross-subject multimodal emotion recognition based on hybrid fusion. IEEE Access 8:168865–168878
15. Alam M, Vidyaratne LS, Iftekharuddin KM (2018) Sparse simultaneous recurrent deep learning for robust facial expression recognition. IEEE Trans Neural Netw Learn Syst 29(10):4905–4916
16. Nidhi Verma B (2023) From methods to datasets: a detailed study on facial emotion recognition. Appl Intell 53(24):30219–30249
17. Jarraya SK, Masmoudi M, Hammami M (2020) Compound emotion recognition of autistic children during meltdown crisis based on deep spatio-temporal analysis of facial geometric features. IEEE Access 8:69311–69326
18. Guo J, Zhou S, Wu J, Wan J, Zhu X, Lei Z, Li SZ (2017) Multimodality network with visual and geometrical information for micro emotion recognition. In: 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017), pp 814–819. IEEE
19. Li S, Deng W, Du J (2017) Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2852–2861
20. Kamińska D, Aktas K, Rizhinashvili D, Kuklyanov D, Sham AH, Escalera S, Nasrollahi K, Moeslund TB, Anbarjafari G (2021) Two-stage recognition and beyond for compound facial emotion recognition. Electronics 10(22):2847
21. Shaila S, Shivamma D, Monica U, Tejashree K (2022) Facial expression recognition for compound emotions using mobile net architecture. In: 2022 International conference on artificial intelligence and data engineering (AIDE), pp 187–190. IEEE
22. Liang L, Lang C, Li Y, Feng S, Zhao J (2020) Fine-grained facial expression recognition in the wild. IEEE Trans Inf Forensic Secur 16:482–494

23. Pons G, Masip D (2020) Multitask, multilabel, and multidomain learning with convolutional networks for emotion recognition. IEEE Trans Cybern 52(6):4764–4771
24. Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y et al (2023) Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 4015–4026
25. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: International conference on learning representations. https://openreview.net/forum?id=YicbFdNTTy
26. Ma F, Sun B, Li S (2021) Facial expression recognition with visual transformers and attentional selective fusion. IEEE Transactions on Affective Computing
27. Lian Z, Sun L, Sun H, Chen K, Wen Z, Gu H, Liu B, Tao J (2024) Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. Inf Fusion 108:102367
28. Zhu A, Li K, Wu T, Zhao P, Hong B (2024) Cross-task multi-branch vision transformer for facial expression and mask wearing classification. J Comput Technol Appl Math 1(1):46–53. https://doi.org/10.5281/zenodo.11083875
29. Dong R, Lam K-M (2024) Bi-center loss for compound facial expression recognition. IEEE Signal Processing Letters
30. Xue F, Wang Q, Guo G (2021) Transfer: Learning relation-aware facial expression representations with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3601–3610
31. Zhao X, Lv Y, Huang Z (2022) Multimodal fusion-based swin transformer for facial recognition micro-expression recognition. In: 2022 IEEE international conference on mechatronics and automation (ICMA), pp 780–785. https://doi.org/10.1109/ICMA54519.2022.9856162
32. Xue F, Tan Z, Zhu Y, Ma Z, Guo G (2022) Coarse-to-fine cascaded networks with smooth predicting for video facial expression recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2412–2418
33. DeVries T, Taylor GW (2017) Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552
34. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2018) mixup: Beyond empirical risk minimization. In: International conference on learning representations, pp 1–13
35. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y (2019) Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6023–6032
36. Jiang J, Wang M, Xiao B, Hu J, Deng W (2024) Joint recognition of basic and compound facial expressions by mining latent soft labels. Pattern Recognit 148:110173
37. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
38. Kim K, Ji B, Yoon D, Hwang S (2021) Self-knowledge distillation with progressive refinement of targets. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6567–6576
39. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2022) Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16000–16009
40. Dai Z, Liu H, Le QV, Tan M (2021) Coatnet: Marrying convolution and attention for all data sizes. Adv Neural Inf Process Syst 34:3965–3977
41. Xia Y, Yu H, Wang X, Jian M, Wang F-Y (2021) Relation-aware facial expression recognition. IEEE Trans Cognit Dev Syst 14(3):1143–1154
42. Vo T-H, Lee G-S, Yang H-J, Kim S-H (2020) Pyramid with super resolution for in-the-wild facial expression recognition. IEEE Access 8:131988–132001
43. Kollias D (2023) Multi-label compound expression recognition: C-expr database & network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5589–5598
44. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
45. Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y (2022) Coca: Contrastive captioners are image-text foundation models. Trans Mach Learn Res 2022
46. Srivastava S, Sharma G (2024) Omnivec: Learning robust representations with cross modal sharing. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1236–1248

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Nidhi received a Bachelor of Technology and a Master of Technology in Computer Science from Banasthali Vidyapith, Rajasthan, India. She is currently pursuing her Ph.D. degree from Delhi Technological University, Delhi, India. Her current research interest is facial emotion recognition.

Bindu Verma is an Assistant Professor in the Department of Information Technology at Delhi Technological University(Formerly known as Delhi College of Engineering). She earned master's and Doctorate degree in Automated Intent Recognition using Hand Gesture and Face Expression Analysis from the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi. She is passionate about working in the area of computer vision, machine learning, human-computer interaction, intelligent systems, and affective state monitoring. I have made substantial contributions to the field of human-computer interaction, emotion recognition, and intent recognition with many research articles published in international conferences and journals. I am the reviewer of many International Journals such as IET Intelligent Transport System, IEEE Transactions on Circuits and Systems for Video Technology, Intelligent Transportation System conferences, etc.

International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-4, April 2020

A Machine Learning and Augmented Reality based Framework for Multilingual Product Identification in Retail using Mobilenets and Vuforia

Geetanjali Bhola, Amogh Bansal, Divij Aggarwal, Gagan Kishor Upadhyay

Abstract*: *In the modern world, Machine Learning and Augmented Reality have taken the retail industry by storm. Machine Learning and Augmented Reality have provided a major boost to the industry of interactive retail by providing features such as real-time product detection and identification. The proposed research aims at overcoming several challenges in the present scenario which include the time consuming process of standing in long queues while purchasing the products at supermarkets, personalizing the shopping experience in order to maintain the privacy of the users, helping the customers to maintain their specified budgets, reducing the high labor costs and overcoming the language barriers while pertaining to selling products to groups with different linguistic backgrounds by combining the real-world interaction of Augmented Reality and Machine Learning-based product identification. This proposed research work aims at providing the customers with a futuristic shopping experience while maintaining their specified budgets. The Machine Learning-based object detection approach detected the products with around 96% accuracy and the Vuforia-based Augmented Reality approach detected objects with maximum accuracy.

Keywords*: *Augmented Reality, Vuforia, Machine Learning, MobileNets, Single Shot Multi-Box Detector.

I. INTRODUCTION

Augmented Reality and Machine Learning have created quite a buzz in the retail industry. Retail in today's economy can be combined with the functionalities of object detection and Augmented Reality to improve the customer experience.

Revised Manuscript Received on April 25, 2020.
* Correspondence Author
Ms. Geetanjali Bhola*, Department of Information Technology, Delhi Technological University, New Delhi, India. Email: geetanjali.getz@gmail.com
Amogh Bansal, Department of Information Technology, Delhi Technological University, New Delhi, India. Email: amoghdtu@gmail.com
Divij Aggarwal, Department of Information Technology, Delhi Technological University, New Delhi, India. Email: divij.aggarwal1410@gmail.com
Gagan Kishor Upadhyay, Department of Information Technology, Delhi Technological University, New Delhi, India. Email: gagankishor98@gmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Nowadays, customers tend to use high-end retail services but the retail industry presently faces a number of obstacles such as reducing the labor costs and eliminating the process of standing in long queues while purchasing various products.
One of the major challenges faced by the retail industry is the language barrier across different regions while maintaining budget constraints which makes shopping across borders quite a tedious task. This research work aims at providing retail services using the real-world interaction of Marker-less Augmented Reality and Machine Learning-based object detection. This provides users with a personalized multi-lingual shopping functionality so as to maintain the privacy of the users [1] with features such as accurate real-time product detection, display of nutritional contents, product characteristics and video advertisements while maintaining the specified budget constraints, presentation of purchase history using login credentials and SQL databases, all available in multiple languages across different regions. The ability to allow the customers to view the cart and instructions on how to use the product in their local language improves the overall shopping experience of the customer [2]. Object detection helps to locate certain classes in the image. Augmented Reality allows coupling and superimposition of computer-generated imagery with real-world entities [3].
AR also helps to improve the user interface (UI) of the applications, improve the speed of local target recognition, use of technology that provides highly robust target-tracking of low-light and partly covered targets while pertaining to language barriers [4]

II. CONCEPTUAL FRAMEWORK

A.Object detection using TensorFlow

The first step involved use of TensorFlow for Object Detection of products in case of the Machine Learning based approach. The TensorFlow Object Detection API is an open source framework which simplifies the detection of objects and makes construction, training and deployment of object detection models easy as it uses pretrained object detection models. It also acts as a medium for detecting objects and algorithmic implementation. TensorFlow is given the name because it takes multidimensional arrays, also known as tensors as input to the model. It can be used with small changes on the several diverse frameworks.

© Copyright: All rights reserved.

The framework is flexible in the sense that it is able to adjust to new conditions and hence, can be used for expressing several different algorithms.

B. Object detection using MobileNets and SSD

MobileNet [5] is based on depth wise convolution in the case of which each filter contains a single input. A light MobileNet based model must include at least five layers with feature size 14×14×512. There can be 2 types of MobileNets design models, thinner or shallower in which the thinner model is more efficient than the shallow model.

Before the SSD, several different algorithms for object detection like R-CNN, Fast- RCNN etc. were obtained by us. R-CNN [6] works by suggesting different spots present in an image that act as input to an object classifier at the time of training. Fast- RCNN [7] and Faster-RCNN [8] are the improved versions of R-CNN, they use some techniques so as to reduce the manipulation of proposed regions. The problem of inference speed for real life objects was overcome by the YOLO [9] system. YOLO was not able to identify smaller objects in the frame, which was its main limitation. The problem was solved by SSD [10], the mAP value approached 74 percent and detection speed was similar to that of YOLO [11].

SSD serves the purpose of real time detection of various objects. Faster R-CNN makes use of a region proposal network which creates boundary boxes so that the objects belonging to several different classes can be detected whereas in case of SSD, region proposal network is no longer needed as it helps to fasten the process and hence it is eliminated. To regain the accuracy, SSD makes minor technical advancements including the default boxes and multi-scale features, using which, SSD is capable of meeting the accuracy achieved by Faster R-CNN with the use of low-resolution object images. The working of SSD was divided in two parts i.e. extracting the feature maps and application of convolution filters. SSD makes the use of small convolution filters to compute both the class and location scores. After the above steps, in order to make predictions about the classes of the detected objects, feature maps are extracted and 3×3 convolution filters are applied on all cells individually. MobileNets and SSD provide the benefit of achieving lag-free accurate object detection on a system with nominal specifications.

C. Creation of SQLite Database in Machine Learning

The next step involved the creation of SQLite Database which serves as an open-source SQL database engine that doesn't require a separate server to operate. It is an in-process library that implements a self-contained, zero-configuration, serverless, transactional SQL database

Engine that has bindings to several programming languages such as Java, Python, C++, Basic [12]. In our case we have made use of Python so as to create a database in case of the Machine Learning based approach which serves the purpose of storing the login credentials as well as the purchase history pertaining to the previous visits of the users.

D. Vuforia based Image Recognition and superimposition of virtual information

In order to detect products in Augmented Reality, we make use of Vuforia which is an AR software development kit which was developed by Qualcomm and allows the creation of various Augmented Reality applications. It allows the recognition and tracking of planar images and 3-D objects in real-time. Virtual objects are placed through a camera viewfinder and then the position of objects is adjusted in the background of the objects and in front of the lens.

In the development process, Vuforia uses target images but any kind of image does not qualify as a target image. All the uploaded images, at first, are evaluated by the Vuforia web developer. Scoring of value target defines how well an image can be detected and tracked using Vuforia SDK. The score of an image acts as a target range of 0 to 5 which helps in assessment [13]. Greater number of stars allows to obtain stronger detection and tracking capabilities of an image target.

In order to determine the number of stars, various features serve as the attributes of the assessment criteria. Repetitive patterns should not be there in the image target and the image target should contain high local contrast, uniform distribution of features and high feature density, so as to obtain better quality of recognition of the image.

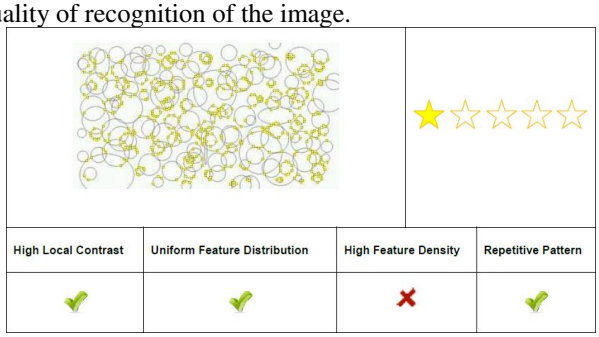


Fig. 1. IT with 1-star ball shape

For example, Fig. 1 which has been shown above, has a quality level in one star because the density of the features is poor. Hence, alteration of its line shape or texture is required, so as to increase (multiply) the number of features of this figure [14].

Fig. 2. IT with 4 stars square shape

Now, Fig. 2, which has been shown above, depicts the rise of quality in star level of the image target by increasing the number of features to improve feature density and similarly for different image targets the quality in star level can be improved by also increasing the local contrast, distributing the features uniformly and through avoidance of repetitive patterns.

This information was integrated in the research work by implementing Marker-Less Augmented Reality which helped in obtaining stronger detection and tracking capabilities of the image targets.

© Copyright: All rights reserved.

The ability of augmented reality technology to blend real scenes and virtual information was used for the superimposition of nutrient distribution and advertisement over a particular product along with information about the remaining budget and the shopping cart.

E. Database Creation using MAMP in AR

Creation of a database in the Augmented Reality approach was carried out using MAMP (Macintosh, Apache, MySQL, and PHP) which allows you to have access to a local PHP server and MySQL server and provides you with all the tools necessary to run WordPress on your machine, for testing and development purposes. The created database allowed storage of the login credentials of the different users in the Augmented Reality-based approach.

III. METHODOLOGY

The main aim of the research was to provide consumers with an interactive experience through real-time product identification in the retail sector.

The flow of the implementation of this idea has been shown in Fig. 3. At first, object detection was carried out with the help of TensorFlow, which because of its high GPU intensiveness and low efficiency resulted in low accuracy and fps, along with lag in product detection. So as to overcome these problems, product detection was therefore implemented using a joint version of Single Shot MultiBox Detector (SSD) structure and Mobile Net classifier [15].

On observation of improvement in detection accuracy and decrease in lag in the latter method, new features were added to enhance the overall user experience. Features such as advertisement, description of the products and interactive budget management for customers were added. Then, a database was created using Python so as to store the login credentials of various users and also provisions for different languages were provided in case of the Machine Learning-based approach. Multiple windows were being required for different features which posed to be a limitation of the Machine Learning-based approach,hence the focus was shifted towards incorporation of Augmented Reality in the research.

At first, the implementation of marker-based AR was done, which performed real-time superimposition of nutritional contents on their respective markers for various products.The marker-less technique was executed due to the additional cost of production of markers, and using marker-less technique, real-time superimposition was done without the use of predefined markers for various products.

Then in order to provide a better customer experience, new features such as display of advertisements, nutrition pie charts of different products as well as interactive budget management with shopping-mart guide for users. Finally, a database was created using MAMP so as to store the login credentials of the users and the option of choosing between various languages was provided.

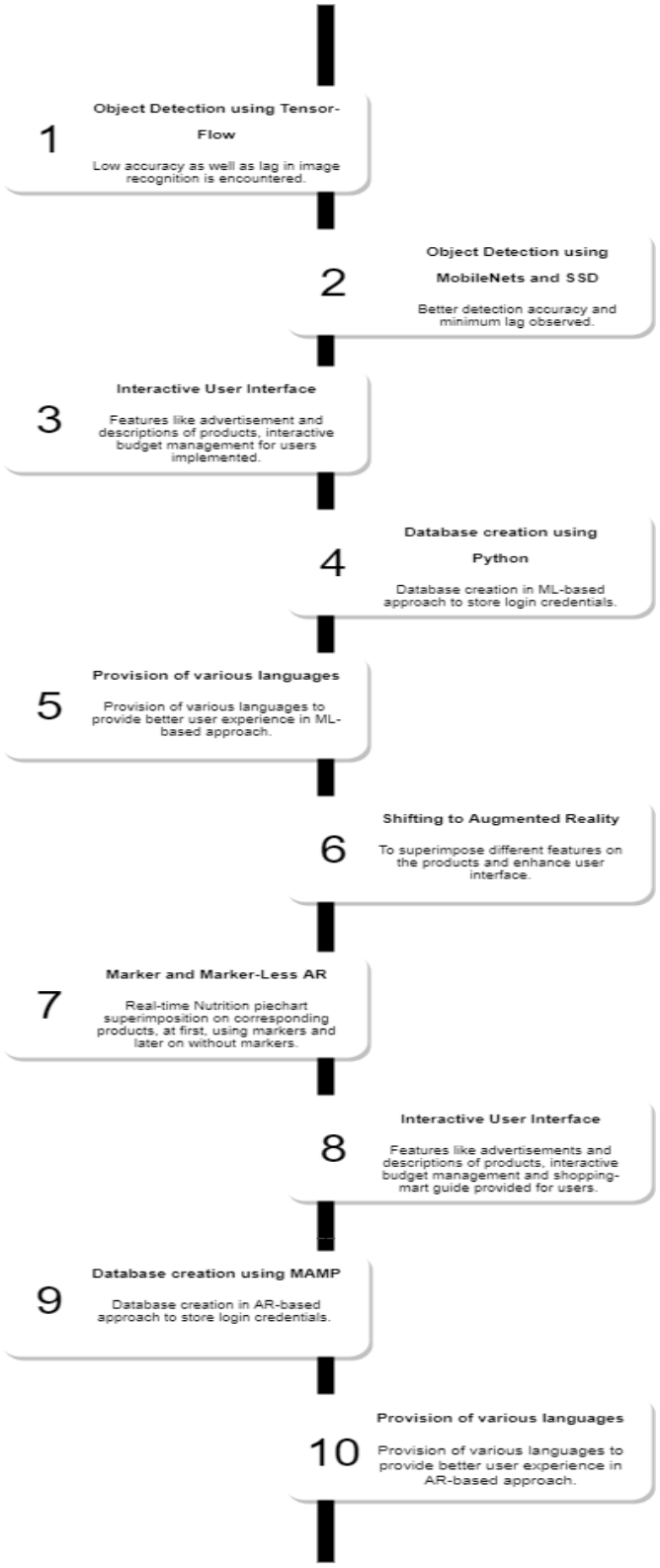


Fig. 3. Project Flow

A. Technology Stack

1. OpenCV

OpenCV is a large open-source library for Machine Learning, Computer Vision as well as Image Processing. It plays a great role in real-time operations. It performed the task of processing images and videos to detect objects and products.

© Copyright: All rights reserved.

OpenCV was the first piece of technology used to detect objects using Python in case of the Machine-Learning based approach.

2. AR Camera

In the case of Augmented Reality, AR Camera was needed to be imported in unity which is used to provide AR scenes. It takes the help of WebCamBehaviour class, VuforiaBehaviour class, DatabaseLoadBehaviour class to accomplish its functions. The WebCamBehaviour class, which also serves as the prime camera control class, informs about whether to switch on the main camera, which camera to use etc. The control entry of Vuforia is the VuforiaBehaviour class, which registers all the open interfaces of Vuforia. The DatabaseLoadBehaviour class is used to trigger identification library, which can perform the function of loading multiple databases that have been uploaded to Vuforia for processing and specifying which library to trigger [13].

B. Implementation

1. Coco Dataset for mobilenets and SSD based object detection

Common Objects in Context is a large-scale object detection, segmentation, and captioning dataset. COCO includes several features such as object segmentation, recognition in context, 330K images and 1.5 million object instances.

We have created a dataset similar to COCO dataset [16], for performing the Machine Learning based Product Identification, that has images of shopping objects like bottles and chairs. Nearly 300 images were taken of both the objects, amongst those images, 60% were utilized in the training phase and 40% in the testing phase. On the detection of multiple objects by the web camera, then a bounding box is drawn over every single detected object with its class as the heading of the bounding box.

2. Database Creation for image targets (Quality of Image targets)

The next step was the construction of the database of image targets for Augmented Reality based Object Detection using Vuforia SDK following which the license key of the created database was added to the research work. At first, the database consisted of the stickers which were used as image targets for products in marker-based object detection in AR. Since the initial rating of the stickers was low (as mentioned in Table 1), hence weaker detection and tracking ability of the image were experienced. Then so as to eliminate the concept of stickers and enhance the trackability of the image, Marker-Less product identification was implemented in AR. The database that was prepared for the marker-less execution consisted of images of actual products which were used as image targets for superimposition. The rating of both the image targets improved significantly (Refer Table 1) and reached 5 stars that ensured better trackability and detection of the images and hence enhanced the functioning of the model.

Table I- Target Score

No.	*Target Image (Marker-Based)*	*Star*	*No.*	*Target Image (Marker-Less)*	*Star*
1	Apple Sticker	2	1	Knorr Soup Image	5
2	Maggie Sticker	2	2	ACT II Popcorn Image	5
	Average (Marker-Based)	***2***		***Average (Marker-Less)***	***5***

3. Provision of Optional Languages

The option of choosing from various languages was then provided in both the Machine Learning and Augmented Reality based approaches in order to overcome the challenges of language barriers across different regions as per the requirement of the users.

4. Database Creation for purchase history and Login Credentials

Finally, databases were created for storing the purchase history as well as the login credentials of the users. In case of the Machine Learning-based approach the SQLite database [12] was created using Python and in case of the Augmented Reality-based approach a database was created using MAMP so as to store the login credentials and the purchase history of the users.

IV. RESULTS

We have tested the machine learning part of our research on an i5 processor laptop. In the machine learning based implementation of the research work, at first, the object detection is carried out using TensorFlow (Fig. 4), but due low accuracy as well as lag in image recognition, the focus was shifted towards object detection using MobileNets and SSD (Fig. 5). After that, the user is made to login (Fig. 6) which provides the user an option to view the user's purchase history (Fig. 7) and then the user can choose a language of his/her choice (Fig. 8). Then, the user enters the budget for shopping (Fig. 9) and after the budget is provided, the process of object detection starts. Once the object is detected successfully (Fig. 10), then the user is asked whether he/she is interested to view the description of the product (Fig. 11) or viewing the advertisement associated with the product. After the user makes the choice, the user has the option to either buy it or ignore it. If the user decides to buy the product, then the item is added to the cart. The user can access the cart anytime to know about the existing products in the cart and the total amount. However, when a user exceeds the budget entered by the user initially, a pop-up showing OVER-BUDGET notification appears and further purchasing of items can't be performed (Fig. 12). Hence, this whole system allows the user to automatically detect any product of interest and then take the decision of buying it or not after seeing the advertisement and nutritional contents of the product.

© Copyright: All rights reserved.

Fig. 4. Object detection using TensorFlow

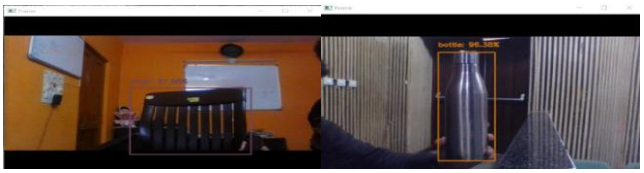


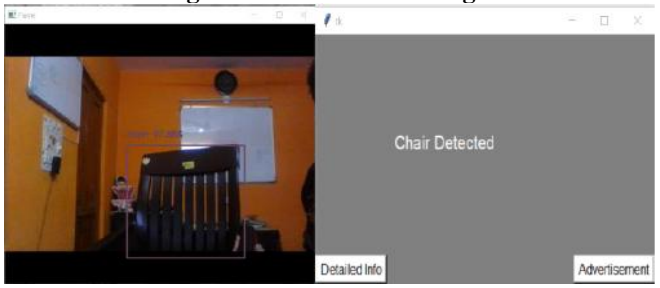
Fig. 5. Object detection using MobileNets and SSD

Fig. 6. User is logged in

Fig. 7. User can view Purchase History

Fig. 8. User can choose language

Fig. 9. User enters the budget



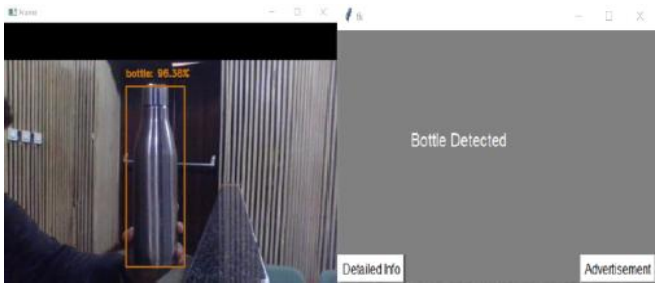


Fig. 10. Product detected

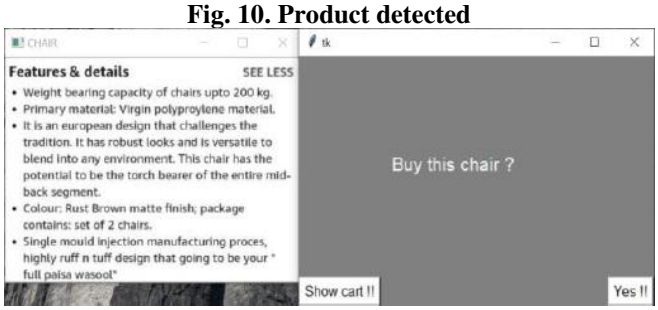


Fig. 11. User can view description

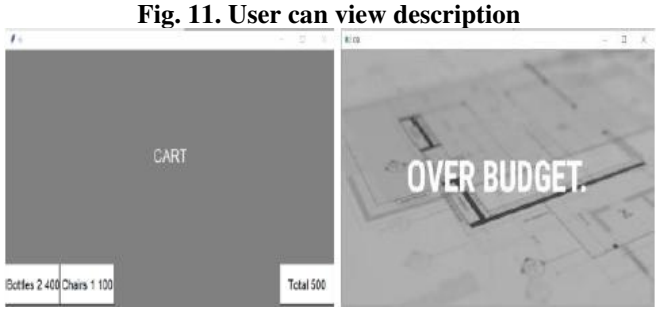


Fig. 12. Over Budget

The main limitation of this approach was that there were several pop-up windows through which the user had to go through to reach the final step, and hence, we then decided to perform the same implementation in Augmented Reality, where the user could interact with the features easily without switching between several windows.

In the marker-less AR implementation of the research work, initially the user is not registered (Fig. 13). Once the user completes the registration process (Fig. 14), his/her login credentials are visible in the database (Fig. 15). After that the user can sign into the login portal using his/her login credentials (Fig. 16). Once the user was logged in (Fig. 17), the product was scanned by the Vuforia SDK and the UI consisting of panels and shopping-mart guide was superimposed on the product (Fig. 18). After that, the user is made to enter the budget for shopping which is then displayed on the panel superimposed at the top of the product. The panel which is superimposed on the left side of the product shows the various choices for the users which include choosing the language preferred by the user, display of advertisement and nutrition value of the product along with add to cart and show cart options (Fig. 19). After viewing the advertisement and the nutrient value of the product, the user has the option to buy the product or not. If the user chooses to buy it, then the product is added to the cart and the budget is updated. The user has also been provided with an option to view the cart and the remaining budget (Fig. 20). Finally, the user can also view his/her purchase history till date (Fig. 21). Hence, this whole model provides an opportunity for users to select the products which are in their budget as well as are perfectly suited for their requirements.

© Copyright: All rights reserved.

Fig. 13. User is not registered

Fig. 14. Registration process

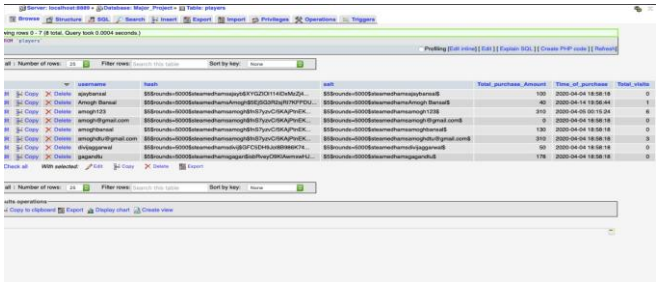


Fig. 15. Login credentials of the user are added to the database

Fig. 16. User logs in using his/her login credentials

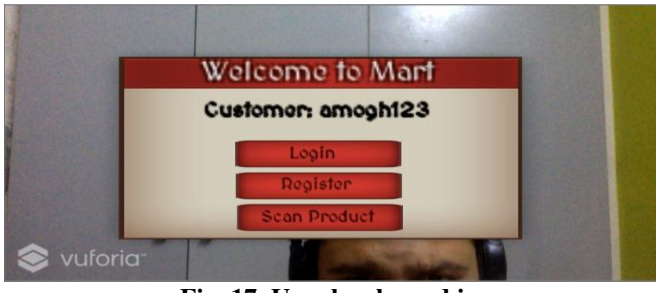


Fig. 17. User has logged in

Fig. 18. Product has been scanned

Fig. 19. Product scan with language change, budget entry, language entry, video and nutrition pie-chart display

Fig. 20. Product scan with show cart and remaining budget

Fig. 21. Purchase history

The summarized test cases of the important functions of the AR (Marker-Less) implementation of the research work is shown in Table 2.

Table II- Test cases for important functions

No.	Precondition	Operation	Expected Results	Tested Results
1.	User not logged in	"Register button is clicked"	User is redirected to the registration portal	Pass

© Copyright: All rights reserved.

2.	Login Credentials of new users are added	"Register" button is clicked	Registration of new user complete, login credentials are added to database and the user is redirected to the login portal	Pass
3.	User Registered	Login Credentials of existing users are added	Login Credentials are displayed in text format on the panel	Pass
4.	Login Credentials of existing users are added	"Login" button is clicked	User is logged in	Pass
5.	User is logged in	"Scan Product" button is clicked	Product is scanned	Pass
6.	Product is scanned	Language of user's choice is entered	Changes displayed according to the language entered	Pass
7.	Product is scanned	Click on the shopping mart guide	"Hi! Welcome to mart" text is displayed and waving animation is performed by the guide	Pass
8.	Product is scanned	"Advertisement" button is clicked	Product advertisement with audio is shown in 3-D format	Pass
9.	"Advertisement" button is clicked	"Advertisement" button is clicked again	3-D advertisement is removed	Pass
10.	Product is scanned	"Nutrition" button is clicked	Product's nutritional pie chart with audio is shown in 3-D format	Pass
11.	"Nutrition" button is clicked	"Nutrition" button is clicked again	3-D display of product's nutritional pie chart is removed	Pass
12.	Budget is entered	Enter Budget for shopping	Budget is displayed in text format on the panel	Pass
13.	Budget is entered	"Add to Cart" button is clicked	Product is added to the cart and the updated budget (Current budget – price of the product scanned) is displayed	Pass
14.	Product is added to the cart	Wait for a few seconds	Shopping mart guide displays "Thank you for choosing (product name)" in text format on the panel	Pass
15.	Products are added to the cart	"Show Cart" button is clicked	Products in the cart are displayed in text format on the panel	Pass
16.	Budget becomes zero	"Add to Cart" button is clicked	"Over Budget" is displayed in text format on the panel	Pass
17.	Budget becomes zero	"Add to Cart" button is clicked	Shopping mart guide displays "Please check your budget" in text format on the panel	Pass
18.	Shopping mart guide is clicked	Shopping mart guide is clicked again	Text displayed by shopping mart guide is removed	Pass
19.	Product is scanned	"Purchase History" button is clicked	The entire purchase history of the customer including the customer's name, total amount spent till date, date and time of last purchase and total number of visits till date are displayed in text format on the panel	Pass
20.	At least 1 product is scanned	Scan another product	Features like budget and cart content are transferred to the newly scanned product's user interface (UI)	Pass

V.CONCLUSION AND FUTURE WORK

The research provides the retail functionality combined with the real-world interaction of Augmented Reality and Machine Learning-based product identification, thus providing a fresh experience to the customers. Supermarket shopping is provided in a stimulating and interesting manner to the customers.

Hence, the research work's design goals were achieved and therefore, research work's future vision includes increasing the scalability of the research work to include a larger database of users, provisions for more languages, a wide range of products in various industries such as sports, footwear, groceries and clothing Also, future work includes implementation of a recommender system for providing product suggestions to customers.

REFERENCES

1. A. Poushneh, "Augmented reality in retail: A trade-off between user's control of access to personal information and augmentation quality," Journal of Retailing and Consumer Services, vol. 41, pp. 169-176, 2018.
2. A. Poushneh and A. Z. Vasquez-Parraga, "Discernible impact of augmented reality on retail customer's experience, satisfaction and willingness to buy," Journal of Retailing and Consumer Services, vol. 34, pp. 229-234, 2017.
3. J. Carmigniani and B. Furht, "Augmented reality: an overview," in Handbook of augmented reality, Springer, 2011, pp. 3-46.
4. F. Peng and J. Zhai, "A mobile augmented reality system for exhibition hall based on Vuforia," 2017.
5. A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," 2017.
6. R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014.
7. R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015.
8. S. Ren, K. He, R. Girshick and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015.
9. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016.
10. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu and A. Berg, "SSD: Single Shot MultiBox Detector," 2016.
11. S. Kanimozhi, G. Gayathri and T. Mala, "Multiple Real-time object identification using Single shot Multi-Box detection," 2019.
12. S. T. Bhosale, T. Patil and P. Patil, "Sqlite: Light database system," International Journal of Computer Science and Mobile Computing, vol. 4, no. 4, p. 882, 2015.
13. D. Adrianto, M. Hidajat and V. Yesmaya, "Augmented reality using Vuforia for marketing residence," 2016.

© Copyright: All rights reserved.

14. A. Simonetti Ibañez and J. Paredes Figueras, "Vuforia v1. 5 SDK: Analysis and evaluation of capabilities," 2013.
15. Y. Wang, C. Wang and H. Zhang, "Combining a single shot multibox detector with transfer learning for ship detection using sentinel-1 SAR images," Remote sensing letters, vol. 9, no. 8, pp. 780-788, 2018.
16. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. Zitnick, "Microsoft COCO: Common Objects in Context," 2014.

AUTHORS PROFILE

Ms. Geetanjali Bhola is an Assistant Professor in Information Technology Department of DTU. She has been taking subjects like database, computer graphics, programming etc. She has co-authored many research papers in the field of image processing. With a deterministic attitude she has helped and encouraged her students to take up challenging research problems which have later won awards at different platforms. Her current research includes augmented reality applications, image processing to name a few.

Amogh Bansal is pursuing B. Tech. in Information Technology from Delhi Technological University, Delhi, India. He has published a research paper titled 'Augmented Reality and Machine Learning based Product Identification in Retail using Vuforia and MobileNets' and presented at the International Conference on Inventive Computation Technologies (ICICT-2020) held at Coimbatore, India from 23-26th February, 2020. He is also the Corporate Head of the Society of Robotics of Delhi Technological University (SRDTU).

Divij Aggarwal is pursuing B. Tech. in Information Technology from Delhi Technological University, Delhi, India. He has published a research paper titled 'Augmented Reality and Machine Learning based Product Identification in Retail using Vuforia and MobileNets' and presented at the International Conference on Inventive Computation Technologies (ICICT-2020) held at Coimbatore, India from 23-26th February, 2020. He has successfully mentored a batch of 50+ students in Data Structures and Algorithms in Java.

Gagan Kishor Upadhyay is pursuing B. Tech. in Information Technology from Delhi Technological University, Delhi, India. He has published a research paper titled 'Augmented Reality and Machine Learning based Product Identification in Retail using Vuforia and MobileNets' and presented at the International Conference on Inventive Computation Technologies (ICICT-2020) held at Coimbatore, India from 23-26th February, 2020. He also has participated and won the I.S.G.W Appathon held in New Delhi, 2017.

© Copyright: All rights reserved.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

A Metaheuristic Approach based Feasibility Assessment & Design of Solar, Wind and Grid Powered Charging of Electric Vehicles

Fayez Alanazi[1], Mohd Bilal[2], Ammar Armghan[3], Senior Member, IEEE, Md Reyaz Hussan[4], Student Member, IEEE

[1]Civil Engineering Department, College of Engineering, Jouf University, Sakaka, 72388
[2]Department of Electrical Engineering, SND College of Engineering and Research Center, Yeola, Nashik-423401
[3]Department of Electrical Engineering. College of Engineering, Jouf University, Sakaka 72388, Saudi Arabia
[4]Department of Electrical Engineering, Aligarh Muslim University, India-202001

Corresponding author: Ammar Armghan (e-mail: aarmghan@ju.edu.sa).

" The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number 223202."

ABSTRACT This study assesses the integration of renewable energy sources with grid-connected systems to power electric vehicle charging stations (EVCSs) in Riyadh, Kingdom of Saudi Arabia (KSA), emphasizing its significance for sustainable development. An environmentally friendly EVCS model and a novel metaheuristic optimization technique, the Giza Pyramids Construction Algorithm (GPCA) is employed, to determine the optimal sizing of system components. The objective is to ensure dependability, as determined by the lack of power supply probability (LPSP), while minimizing the levelized cost of energy (LCOE) and total net present cost (TNPC). We assess the efficacy of other algorithms in scaling system components by contrasting the performance of the GPCA with those of the Artificial Hummingbird Algorithm (AHA), Flower Pollination Algorithm (FPA), and Whale Optimization Algorithm (WOA). The study also examines the impact of climatic, financial, and technical factors on system costs, providing insights into the challenges, opportunities, and socio-economic factors influencing the deployment of EV infrastructure in KSA. Findings suggest that the optimal EVCS configuration is a solar photovoltaic/wind turbine (SPV/WT) grid-tied system, resulting in an LCOE of $0.0796/kWh, a TNPC of $102,042, and an annual CO_2 emission reduction of 18,840 kg. The system achieves a 67.9% renewable penetration rate with 30.8% annual grid electricity purchase and 32.4% sale. Sensitivity analysis on renewable parameters is also conducted, demand charges, grid purchase prices, discount rates, and inflation rates to assess their impact on the system's economic variables. By identifying current constraints and potential strategies, our study aims to inform policy-making and investment decisions for expanding electric vehicle charging infrastructure, with broader implications for developing nations.

INDEX TERMS Electrical vehicle charging stations; Renewable power resources; Kingdom of Saudi Arabia, Sustainable development; Giza Pyramids Construction Algorithm; Lack of power supply probability; CO_2 emission.

I. INTRODUCTION

The implementation of electric vehicles (EVs) signifies a fundamental change in the transportation sector through the provision of cost-effective alternatives and a significant mitigation of greenhouse gas emissions. The potential for these developments to improve public health and alleviate the impacts of climate change is considerable [1]. However, the substantial energy demand associated with charging EVs on a large scale presents considerable challenges [2]. The broad use of EVs is closely tied to the development of robust charging networks, crucial for increasing their appeal and user acceptance. Thus, the growing demand for EVs highlights the necessity to upgrade and expand existing charging infrastructures to facilitate a rapid societal transition towards electric mobility [3]. Given the heightened interest in EVs and the abundant generation of renewable energy in the EV sector,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

it is essential to conduct a comprehensive analysis of the technology, economics, and environmental aspects of EV charging stations utilizing hybrid intermittent generation systems [4]. The merging of Renewable Energy Sources (RESs) with EVs is crucial in attaining sustainable development goals (SDG), notably SDG7 (Ensuring access to affordable, reliable, sustainable, and modern energy for all) and SDG13 (Taking urgent action to combat climate change and its impacts) [5]. Furthermore, an optimal EV charging framework that leverages a combination of renewable sources, including solar photovoltaics (SPV), wind turbines (WT), battery energy storage (BES), grid connections, and various bio-energy options, is acknowledged for its technical, economic, and environmental advantages [6].

The accomplishment of sustainability objectives is enhanced by merging EVs with the power grid and different RESs using Vehicle-to-Grid (V2G) technology, forming a cooperative hybrid system. V2G's ability to provide electricity back to the grid offers an efficient tactic to alleviate the negative impacts of EV charging requirements [7]. EVs equipped with V2G not only tackle pressing environmental issues like air pollution but also reduce reliance on fossil fuels. Research conducted by [8] explores the economic operations of EV parking lots equipped with V2G-enabled charging facilities, RESs, and Energy Storage Systems (ESS), proposing an optimization model to maximize profits generated with the help of EV charging services and energy contributions to the grid while considering uncertainties, ESS wear, and variable EV parking behaviors. As EVs become more prevalent, they exert significant pressure on electrical grids, particularly when meeting the fluctuating demands of EV charging becomes challenging. The authors analyze the V2G system within Indonesia's JAMALI grid, investigating the effects of various feed-in tariff schemes [9]. Currently, V2G controls in microgrids suffer from inadequate coordination, leading to issues such as voltage instability and power supply disruptions when integrating EVs into residential or commercial microgrids [10]. To address these challenges, researchers in [11] evaluate microgrids incorporating wind and solar energy generation, battery operations, and Compressed Air Energy Storage (CAES), with a focus on V2G scheduling within the IEEE 33-bus network. V2G systems enable two-way electricity exchange between the EV battery and the grid, opening up new economic opportunities for owners of EVs and utility providers. In the context of smart grids, V2G technology is poised to provide additional grid services such as load balancing, peak shaving, and cost reductions for power generation enhancements [12].

The recent increase in electricity demand in KSA, propelled by economic expansion and population growth, has spurred initiatives advocating for renewable energy and changes in the energy composition. Incorporating EVs into the power system signifies a substantial step towards efficiency, potentially evening out the load pattern on the national grid [13]. However, the rapid growth of the EV market presents challenges, particularly in meeting demand on distribution networks. Studies have investigated the impact of EV deployment on electricity demand and distribution networks, proposing models to generate EV demand profiles based on factors such as driving patterns and charging characteristics [14]. RESs in KSA offer a promising opportunity to establish a renewable-powered grid capable of supporting EV charging while ensuring grid stability. Efforts to integrate renewable energy into the electric grid focus on minimizing CO_2 emissions and maximizing energy production from renewable sources [15]. Proposals for hybrid power systems at EV charging stations, incorporating SPV panels, WT, and fuel cells, aim to design a grid reliant on clean energy [16]. The transition toward smart grids and hybrid power systems marks a transformative phase in energy management systems (EMS), shifting toward decentralized systems integrating large-scale renewable resources [17]. KSA is well-positioned to advance smart energy and communication systems, striving to develop smart hybrid power systems that offer both security and flexibility for technological integration in the future [18].

The present circumstances are acknowledged by our research as a chance to conduct thorough evaluations regarding the dimensions, positioning, and integration of renewable energy sources into the power grid. This provides insightful information regarding the socioeconomic factors, obstacles, and prospects that influence the adoption of electric vehicles and charging stations in Saudi Arabia. The region's plentiful wind and solar resources can be efficiently utilized, and our research advocates for a techno-economic hybrid energy system. Such a system, employing a combination of RES, not only supports environmental sustainability objectives but also contributes to the country's economic and energy diversification strategies.

A. LITERATURE REVIEW

EVCSs are widespread across various locations, resulting in varying charging times for EVs. This variability is further amplified when integrating renewable energy resources into EVCSs. In three distinct regions of India, researchers examined the financial and technical viability of grid and renewable energy systems for environmentally sustainable EVCSs [19]. The study involved the optimization of system component sizing in order to minimize the levelized cost of electricity and total net present cost, while ensuring that the probability of power supply failure remained acceptable. This resulted from the implementation of an innovative optimization algorithm based on metaheuristics. This system aims to reduce dependence on strained grids [20]. With increasing concerns about climate change, EVs have emerged as a potent solution to pollution from fossil fuel transportation. To achieve a carbon-neutral transportation sector, EV batteries should be powered by RES. In [21], the authors devised an energy management strategy (EMS) with the objective of regulating the grid's peak and off-peak power requirements. The dimensions and placement of swift EVCS are critical

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

determinants in the distribution network. An advanced bald eagle search algorithm is suggested by the authors with the following objectives: minimize energy loss in the distribution system, decrease transportation costs to EVCSs, and ascertain the most advantageous locations and capacities for EVCSs. Previous research, such as [22], focused on optimal sizing by integrating EVs into the distribution network. By segmenting the network into communities, strategically allocating photovoltaic systems (PVCS) within each community using the backward-forward sweep (BFS) method, and utilizing particle swarm optimization (PSO) to determine the optimal size and quantity of PVCS in order to minimize power loss and increase voltage, this methodology is implemented. This method's efficacy was validated using an IEEE 33-node test feeder. An additional investigation examines the optimization of charging station locations and capacities in Kota City, Rajasthan, India [23]. Other researchers have looked into the integration of capacitors with EV charging stations in distribution networks, examined how distributed generation affects the system's reliability [24]. In [25], the impact of distributed generation and EV charging stations on system reliability is examined. Site surveys, electric vehicle (EV) density, electricity demand analysis, load modeling, cost analysis, EV battery charging time, and power quality were integrated into the development of a mathematical model. Optimized distribution systems for integrating diverse RES, including SPV and WT units, with extant EVCSs in particular locations are the subject of this research [26]. Battery Energy Storage Systems (BESS) were deployed at the PV and WT units' sites. The performance of the distribution systems was improved through the utilization of these BES units, which utilized the Gorilla Troop Optimizer (GTO) algorithm to minimize power loss and total voltage deviation (TVD). This algorithm is applicable in optimization scenarios with a single objective or multiple objectives.

In the current environmental context, vehicle-to-grid (V2G) capable EVs present a viable answer to modern problems including air pollution and fossil fuel dependency. Using RESs such as solar systems and strategically planning EV charging and discharging schedules are two ways to address these problems [27]. EVs are gaining importance in the automotive industry due to their reduced CO2 emissions, simplified maintenance requirements, and decreased operating expenses. The increasing prevalence of electric vehicles (EVs) influences distribution network characteristics including power loss, voltage profiles, and harmonic distortion. In addition, the increased peak power demands that occur when EVs are being charged at charging stations present challenges for both EV users and EVCS investors. The integration of RESs at recharge stations is recommended in order to reduce the energy demand on the grid. In addition to maximizing the use of renewable energy and minimizing peak power consumption, energy management techniques (EMS) that regulate the charge and discharging of BES are crucial for achieving these objectives [28]. EV charging and discharging have a big impact on voltage profiles, operating costs, and the stability of electrical networks. When evaluating and integrating V2G and Grid-to-Vehicle technologies, it is recommended to employ a versatile multi-objective optimization approach that considers financial, environmental, and technological factors. Additionally, this approach considers the battery life cycle of electric vehicles (EVs), the charging and discharging schedules of EV users, and their driving patterns [29]. The increased adoption of electric vehicles (EVs) gives rise to sustainability concerns, such as the need to efficiently and sustainably procure electricity from renewable sources, manage grid demand, and construct additional charging stations. In this context, an optimized architecture for a wind-solar hybrid energy charging station is proposed, utilizing the HOMER software. This approach to measurement is adaptable on a global scale. The optimal configuration of this hybrid system utilizes 44.4% wind energy and 55.6% solar energy, producing 843,150 kWh of electricity annually at a cost of $0.064 per kWh [30].

The integration of renewables into EVCSs has demonstrated increased efficiency and grid balancing benefits. However, it's essential to consider the economic implications of connecting or disconnecting the grid from renewable sources. In [31], the authors conducted an exhaustive investigation into the technical and financial benefits of off-grid and grid-connected hybrid energy systems, including battery energy storage BES systems, solar, and wind power, for EV charging loads in the Qassim region of Saudi Arabia. Utilizing a novel optimization method, the Modified Salp Swarm Optimization method (MSSOA), they aimed to reduce energy costs and power outages. Through a comparison of MSSOA to traditional optimization techniques, the method's effectiveness was validated. The technical and financial feasibility of an electric vehicle (EV) charging initiative fueled by renewable energy sources was assessed in Nigeria across six distinct geographic and climatic regions using HOMER Pro® [32]. However, in [33], the authors point out that the widespread use of EVs for transportation raises new demands on utility grids for electricity, and they recommend using EVCSs in off-grid sites to relieve some of the strain on the current grid network. From a techno-economic standpoint, an innovative off-grid plan that combines RES to supply sustainable electricity to EVCSs is evaluated. In [34], an innovative approach is put forth to evaluate the techno-economic viability of retrofitting current petrol stations with EV Supply Equipment (EVSE), taking into account the incorporation of an onsite photovoltaic system for sustainability and the integration of a BES system with EV charging infrastructure to lower grid connection costs. According to [35], charging electric vehicles at home is favored by 90% of EV owners, which makes it a significant concern for grid administrators and car owners alike. The presentation of solutions spans a number of scenarios and includes developments in smart house design, incentives, software optimization for home energy management, and RES

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

and storage. A real-time model for regulated EV load charging was presented by the authors in [36]. A demand-side management (DSM) strategy was implemented in this model by utilizing queuing analysis in conjunction with nonhomogeneous arrival rates and charging service periods. Reducing operational expenses while meeting the charging needs of electric vehicle (EV) users amidst unpredictability is the objective of Distributionally Robust Model Predictive Control (DRMPC) for energy management of a smart EVCS that incorporates PV and energy storage systems, V2G/Vehicle-to-Vehicle (V2V) technologies presented in [37].

The exploration of the energy efficiency relationship between RESs and EVs requires further investigation to bridge existing gaps and assess how EVs can enhance the sustainability of RESs. Addressing various challenges, including the need for rapid EV charging stations, numerous EV technologies have been employed. In [38], to enhance the efficiency of real-time management and strategic capacity allocation, the authors developed a framework for stochastic dynamic simulation modeling pertaining to a regional network of electric vehicle fast-charging stations. The framework is composed of two components: a multinomial logit station choice model and a multi-server queuing model. The latter is utilized to monitor supply and demand dynamics at each station, while the former incorporates charging costs, projected wait times, and detour distances. In [39], a model for a quick EVCS that is linked to the grid is provided, guaranteeing optimal power transmission with the least amount of harmonic current. The model has an integrated solar PV generation system to lessen grid effect, decentralized control for individual vehicle charging, and a converter connecting the grid to a DC bus. Increased EV use could provide power systems with a number of difficulties, including reliability problems. An extensive infrastructure for EV charging must be taken into account when analyzing the effects of EVs on power system reliability. The authors evaluated different charging infrastructures and customer preferences using the Loss of Load Expectation (LOLE) and Loss of Energy Expectation (LOEE) indices in order to comprehend the influence of EVs on power system reliability. When analyzing the effects of diverse charging infrastructure portfolios, two distinct scenarios are considered: a charging infrastructure share determined by consumer preferences and an equal percentage of all charging infrastructure [40]. SPV and BES combined with smart charging can lessen the detrimental effects of EVs on power distribution networks. Their study's objective, with regard to sizing, was to create an on-grid PV-BES system for Brazilian fast EVCS by utilizing modular solutions to boost PV power based on past load demand and utilizing Homer Grid software to analyze grid impact [41]. In [42], a unique planning approach is put out to coordinate the growth of RES with EV charging infrastructure in power networks. In order to comprehend the interactions between EVs and RES, this model takes into account individual EV charging behaviors, EV aggregation, system economic dispatch, and a Model Predictive Control (MPC) learning technique. Environmental concerns and the depletion of fossil fuels are driving a worldwide trend toward the incorporation of photovoltaics (PVs) and electric vehicles (EVs) into the electric power distribution system. Appropriate control design and schedule optimization are required to maintain system stability in light of the challenges presented by the concurrent integration of PVs and EVs into the power system, which include dynamic loading situations, unpredictable power penetration, and charge. An implementation of a solar-powered grid-tied charging station (SGTCS) is proposed, which optimizes electric vehicle (EV) charging while considering seasonal power generation fluctuations and EV charging connectivity through the use of a scheduling technique [43]. Based on day-ahead solar power generation forecasts, an algorithm for the best EV charge scheduling in solar-powered stations is proposed [44]. The objective of the algorithm is to reduce the total cost of charging. To optimize energy efficiency, reduce operating expenses, and gain effective control over load profiles, energy participants must implement coordinated power demand management at the residential or domestic level. In conclusion, a hierarchical coordination structure for managing residential loads is presented, which incorporates PV units, BESs, and EVs. On an actual power distribution network with real demand patterns and dynamics, this framework is evaluated. It facilitates bidirectional power transfer via Vehicle-to-Grid (V2G) operations, manages real-time profiles of residential loads, and coordinates operations via its controller [45].

Infrastructure for charging EVs must be dependable and easily accessible, as the market is predicted to continue growing at a rapid pace in the years to come. This infrastructure needs to support the growing number of EVs while also integrating RESs and enabling efficient data sharing between different stakeholders [46]. This case study-based article's primary goal is to present a comprehensive analysis of the most recent scientific research on various charging tactics. These tactics include renewable and grid-tied charging stations, conductive and inductive charging techniques, and the data required to charge various EV models.

B. KNOWLEDGE GAP

The prevailing studies in the literature predominantly focus on utilizing solar and wind energy to power EV loads, with a narrow scope concentrating exclusively on these renewable resources. The integration of SPV and WT systems is acknowledged as beneficial for the design of EVCS, considering the inherently variable nature of renewable resources. However, there is a notable paucity in research exploring EVCS and RES under diverse meteorological conditions. The literature reveals a limited number of studies that rigorously evaluate both the technical and economic feasibility of grid-tied, renewable energy-based EVCS.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

Furthermore, several studies neglect to consider key financial factors influencing EV charging systems, failing to conduct comprehensive sensitivity analyses of proposed system configurations.

The current research landscape lacks detailed exploration of Saudi Arabia's specific context in developing an EV charging infrastructure that integrates RESs and the electric grid. There's a crucial need for comprehensive analysis focusing on charging methodologies in major urban centers like Riyadh, accompanied by insights into implementation challenges, potential opportunities, socio-economic impacts, and strategic policy development for EVs and their charging infrastructures. This study aims to address these gaps, with a specific focus on Riyadh due to its abundant renewable energy resources, government-led e-mobility initiatives, readiness for swift policy adoption regarding EVs and EVCS, and the financial capacity of its inhabitants to adopt EVs. These factors are expected to drive the growth of a diverse EV charging infrastructure in Riyadh. Additionally, prevailing strategies utilizing metaheuristic algorithms often suffer from complex designs, an abundance of control parameters, and a tendency to converge on local optima. Many studies rely on intricate mathematical models requiring extensive data for effective operation, primarily developed through analytical methodologies. In contrast, metaheuristic approaches offer advantages such as reduced computational time, lower parameter control complexity, and the ability to generate satisfactory solutions. The proposed Giza Pyramid Construction Algorithm (GPCA) addresses these challenges by offering simplicity in implementation and effective convergence.

Furthermore, within the broader context of Riyadh, the growing trend for EVs underscores the need for a foundational framework for EV charging powered by solar and wind energy. Such a model could serve as a blueprint for other Saudi Arabian cities. Therefore, the establishment of a grid-tied EVCS infrastructure based on renewable resources in Riyadh is not only crucial but also symbolic of a broader national shift towards sustainable urban mobility.

C. CONTRIBUTION

On the basis of the findings of a prior literature review that investigated renewable resources data and EV charging patterns in depth, optimization strategies were implemented to develop the ideal model for the Saudi Arabian city of Riyadh. A concise overview of the primary contributions of this research endeavor is as follows:

- To optimize electric vehicle charging stations (EVCSs) powered by hybrid Renewable Energy (RE) in the Kingdom of Saudi Arabia (KSA), the study investigates the implementation of various algorithms. This includes an exhaustive comparison of various algorithms, which sheds light on how their implementation could enhance the design and functionality of EVCSs. With any luck, this analysis will contribute to the development of environmentally conscious, financially viable, and sustainable electric vehicle charging programs in Saudi Arabia, according to the research.
- The study performs a comprehensive technical and financial analysis of grid-connected EVCS and renewable energy systems that have been installed in Riyadh, Saudi Arabia. The utilization of the advanced Giza Pyramids Construction Algorithm (GPCA) is emphasized in order to maximize the Total Net Present Cost (TNPC) and the Levelized Cost of Electricity (LCOE). The study conducts a comparative analysis of the effectiveness of GPCA in relation to several other algorithms, such as the Artificial Hummingbird Algorithm (AHA), Whale Optimization Algorithm (WOA), and Flower Pollination Algorithm (FPA). This comparative analysis considers various system configurations as well as the optimal sizing of renewable energy components implemented into the grid and diesel generators (DG).
- The study conducts a sensitivity analysis in order to assess the influence of fluctuating environmental and economic variables on the TNPC and LCOE. Solar irradiance, wind speed, the discount rate, demand charges, and the inflation rate are some of these variables. Furthermore, the research investigates the effects of different levels of Lack of Power Supply Probability (LPSP) on these economic indicators, providing valuable insights into the feasibility and adaptability of EVCS in diverse circumstances.

II. METHODOLOGY

The methodology of our study focuses on a thorough analysis of a hybrid energy system that is specifically designed to charge EVs, as illustrated in Fig. 1. This system is meticulously designed with several components: SPV arrays and a WT for harnessing renewable energy; a BES unit for storing surplus power; a DG serving as a backup electricity source; an electric grid connection allowing bi-directional flow of energy in and out of the system; and a bi-directional converter facilitating energy conversion between alternating-current (AC) and direct-current (DC) forms. The SPV, BES, and EV charging load are linked to a DC link, while the DG, WT, and grid connection are attached to an AC bus. This configuration enables the system to store excess renewable energy in the BES or export it to the grid. During periods of insufficient power from the SPV and WT, the BES discharges its stored energy to meet the demand, with additional power sourced from the grid or the DG system if necessary. The bi-directional converter plays a crucial role in maintaining energy balance and flow within the system, ensuring that the output from the different components aligns with the requirements of the EV charging infrastructure. This hybrid system aims not only to optimize energy utilization by integrating renewable and conventional power sources but also to ensure a consistent and sustainable power supply for the effective operation of EV charging stations.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

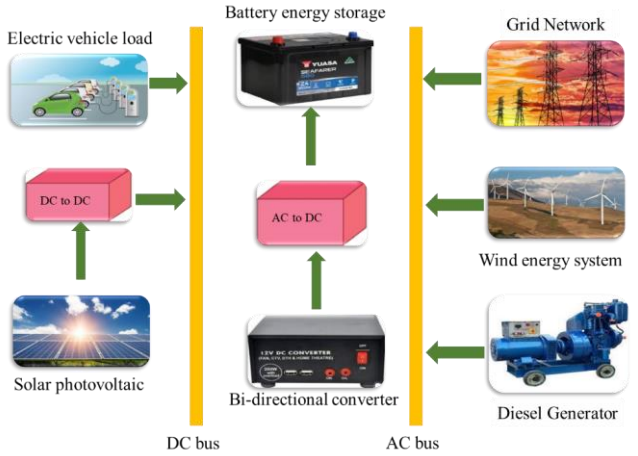


FIGURE 1. **Schematic layout of the proposed hybrid energy system**

FIGURE 2. **Location of the selected city (Riyadh) on the Map**

A. DESCRIPTION OF THE SELECTED SITE AND EV LOAD

Our research focuses on the KSA, the second-largest country in the Arab world, which is located in West Asia and occupies the majority of the Arabian Peninsula, or around 2,150,000 square kilometers (830,000 square miles). The Kingdom of Saudi Arabia is bordered to the north by Jordan, Iraq, and Kuwait; to the west by the Red Sea and Gulf of Aqaba; to the east by the Arabian Gulf, Qatar, and the United Arab Emirates; to the southeast by Oman; and to the south by Yemen. Strategically positioned in the region, KSA encompasses major cities such as Riyadh, the capital, with a population of 6.5 million; Jeddah, a significant sea-port on the Red Sea, with 4.0 million inhabitants; and the religious centers of Makkah (Mecca) and Madinah (Medina), with populations of 2.0 and 1.3 million inhabitants, respectively. Blessed with extensive RES due to its favorable climate, geographic location, and vast, unused landscapes, KSA holds significant potential for harnessing RESs, particularly for fulfilling the charging demands of EVs. The city of Riyadh has been chosen as the focal point for our study. Riyadh stands out for its abundant RESs, its pivotal role as a government hub for large-scale e-mobility development, its potential for rapid adoption of EV-related policies and regulations, and the financial capacity of its residents to afford EVs. This convergence of factors is expected to drive the proliferation of diverse EV charging infrastructure options in Riyadh. The study includes a map illustrating Riyadh's location (Fig. 2), underscoring its strategic importance as a focal point for analyzing the integration of EVCS and the potential for widespread EV adoption within the broader energy and transportation objectives of KSA.

Our research delineates a proposed energy system that has been specifically engineered to fulfill the needs of an EVCS situated in Riyadh, the capital city of the Kingdom of Saudi Arabia. Assuming the installation of this EVCS in Riyadh, the analysis takes into account the charging behaviors and demand patterns of electric vehicles utilizing this station. It is expected that the EVCS will provide service to around 30 to 40 EVs at various time intervals over the course of the day. Based on an analysis of user behavior, it has been determined that approximately one-third of the electric vehicle (EV) owners exhibit a preference for charging their vehicles at night, more precisely from 00:00 to 06:00. This preference is primarily influenced by traffic conditions experienced at the charging station during nighttime hours. Conversely, the remaining two-thirds of EVs are charged during daylight hours at various times. Understanding and managing this distribution of charging patterns are crucial for effectively balancing the load on the energy system. In terms of energy requirements, the EVCS is projected to necessitate approximately 172 kilowatt-hours (kWh) per day, with a maximum demand reaching around 17 kilowatts (kW). This demand profile is pivotal in designing the energy system to ensure it can adequately fulfill the requirements of the EVCS throughout the day and across different seasons. The daily load profile data for EVs at this specific charging station in Riyadh is elaborated in Fig. 3 of the study. The information presented in this data offers significant insights into the station's consumption patterns. These insights form the basis for optimizing the hybrid grid and renewable-powered energy system in order to cater to the station's particular demands. The optimization process holds significant importance in guaranteeing the sustainable and effective functioning of the EVCS, capitalizing on the capabilities of RESs while catering to the varied charging requirements of EV users in Riyadh. The provided data presents valuable insights into the station's energy usage patterns and serves as the foundation for enhancing the hybrid grid and renewable energy system specifically tailored to meet these requirements. It is critical to ensure the efficient and sustainable operation of the EVCS by maximizing the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

potential of RESs and accommodating the diverse charging needs of EV users in Riyadh through this optimization process.

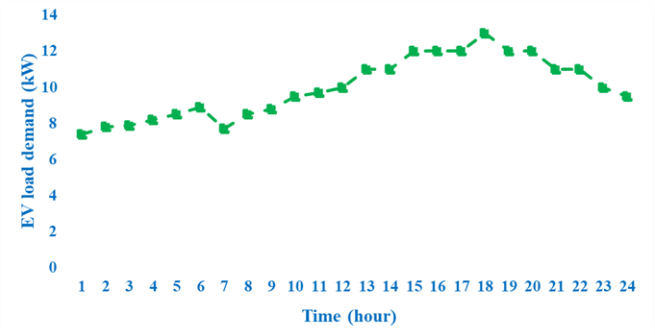


FIGURE 3. Daily EV load profile data of Riyadh city Location of the selected city (Riyadh) on the Map

B. DETAILS OF RESOURCES IN THE SELECTED SITE

Accurate assessment of Renewable Energy (RE) potential holds paramount importance for maximizing power output, particularly in countries like the KSA, which boasts vast land resources suitable for renewable energy installations. This potential plays a pivotal role in achieving KSA's developmental objectives, especially those outlined in its Vision 2030 plan. The efficacy of WT and SPV systems in KSA heavily relies on local wind speeds and solar irradiance. Data on solar irradiance, wind speed, and temperature for the selected site in KSA, obtained from NASA's POWER directory, are elaborated in Fig. 4 and Fig. 5. Fig. 4 illustrates the average monthly solar radiation and wind speed data. It indicates that the peak solar irradiance is recorded in June, while the nadir is observed in December, with an annual average value of 5.84 kilowatt-hours per square meter per day (kWh/m²/day). Regarding wind speed, the highest and lowest values occur in June and October, respectively, with an annual mean wind speed of 5.21 meters per second (m/s). Additionally, Fig. 5 presents details on the average monthly temperature at the study location, highlighting July and August as the months with the highest temperatures, and January and February as the coldest, culminating in an annual average temperature of 29.13°C. KSA's substantial solar potential, evidenced by its global ranking as the seventh among the top ten sites for clean electricity generation through direct sunlight and photovoltaic cells, underscores its capacity to generate significant clean power. These climatic factors – solar irradiance, wind speed, and temperature – are fundamental for strategic planning and implementation of renewable energy projects, particularly in supporting the energy requirements of EV charging stations in Riyadh. This comprehensive comprehension of local RESs is pivotal in developing efficient and sustainable energy solutions that align with KSA's broader environmental and developmental aspirations.

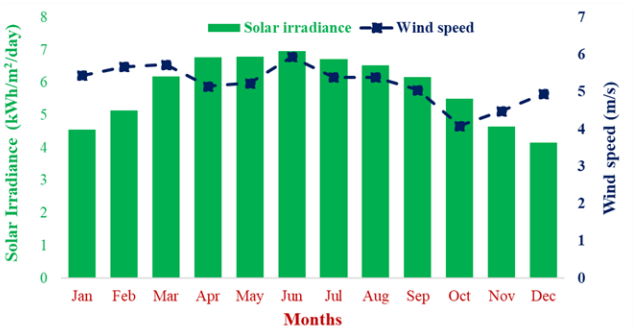


FIGURE 4. Monthly solar radiation and wind speed for Riyadh city

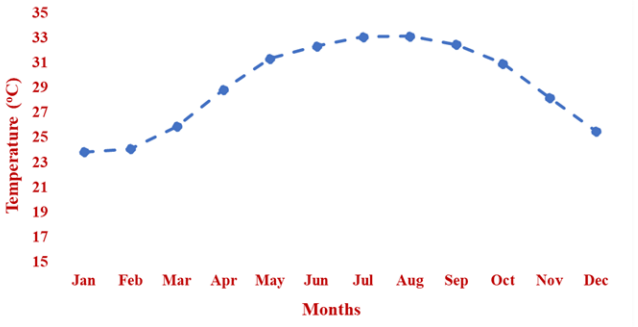


FIGURE 5. Monthly ambient temperature for Riyadh city

C. MODELING OF DIFFERENT COMPONENTS OF INTERCONNECTED ENERGY SYSTEM

The mathematical representation of the energy components within the proposed integrated energy system for recharging EVs is crucial for its layout and operation. The key components of this system, including the inverter, grid connection, DG, WT, SPV unit, and BES system, provided in Table 1, play essential roles in ensuring the overall performance and effectiveness of the hybrid energy system. In the proposed system, the inverter serves to convert the direct current (DC) generated by the SPV unit and stored in the BES into alternating current (AC), aligning it with the grid and DG. The grid connection serves as an additional power source and facilitates the export of excess renewable energy. The DG functions as a backup power source, supplementing the system during periods of insufficient renewable energy generation. The WT and SPV unit harness wind and solar energy, respectively, providing clean and renewable power to the system. A critical component of the system is the BES, particularly the battery, which plays a fundamental role in managing the variability in renewable energy generation. The BES ensures a more stable and reliable power supply for the EV charging requirements. By storing excess energy generated during peak renewable production periods, the BES can release this stored energy during periods of low renewable generation or high EV charging demand. The mathematical modeling of these components involves analyzing their performance characteristics, energy conversion efficiencies, storage capacities, and integration within the overall system. This analysis is essential for understanding the dynamics of

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

energy flow within the hybrid system and for optimizing the system's design to efficiently and sustainably meet the specific energy needs of EVCS. Additionally, the modeling includes considerations for the scalability and adaptability of the system to accommodate varying levels of EV charging demand and changes in renewable energy availability. This ensures that the integrated energy system remains flexible and capable of meeting evolving energy requirements over time.

TABLE 1. Specification of various hybrid system components [47]

Components	Characteristics	Values
Solar PV panel	Rated power	325 W
	Capital cost	1300 $/kW
	Cost of replacement	900 $/kW
	O and M cost	20 $/kW
	Life span	25 years
	Derating factor	80 %
	Module efficiency	16.94%
Wind turbine	Rated power	1 kW
	Capital cost	4450 $
	Cost of replacement	4450 $
	O and M cost	110 $
	Life span	20 years
Diesel generator	Minimum load ratio	50 %
	Operating time	15000 hours
	Capital cost	370 $/kW
	Cost of replacement	296 $/kW
	O and M cost	0.05 $
Battery	Battery material	Lead acid
	Nominal voltage	12 V
	Life span	12 years
	Capital cost	1100 $/kW
	Cost of replacement	1000 $/kW
	O and M cost	10 $/kW/year
Bi-directional converter	Rated power	60 kW
	Capital cost	800 $/kW
	Cost of replacement	750 $/kW
	O and M cost	8 $/kW/year
	Life span	15 years

1) MATHEMATICAL MODELING OF SOLAR PHOTOVOLTAIC ARRAY

The performance and productivity of a SPV system depend on several factors, such as the dimensions of the panels, the amount of sunlight received daily, the temperature of the solar cells, and the geographical location's coordinates. These elements collectively affect the SPV system's ability to convert solar energy into usable electrical power. To measure the power generated by an SPV system accurately, a mathematical equation can be utilized, as outlined in Eqn. (1) [47].

$$\mathrm{P_{SPV}(t)} = \eta_{\mathrm{SPV}} P_{SPV}^{max} \mathrm{n_{SPV}} \eta_{\mathrm{wire}} \frac{I(t)}{1000}\left(1 - \lambda_T\left(T_{amb} + \frac{(NOCT-20)}{800} I_{amb}(t) - 25\right)\right) \quad (1)$$

Where $\mathrm{n_{SPV}}$ denotes the number of SPV cells, P_{SPV}^{max} denotes the maximum power obtained from the SPV cells, η_{SPV} and η_{wire} are the efficiency and wiring efficiency of SPV cells, $\mathrm{P_{SPV}(t)}$ is the power output of the SPV cells, λ_T is the thermal coefficient of the SPV panels, T_{amb} stands for the ambient temperature, $I(t)$ represents the incident solar radiation and $I_{amb}(t)$ denotes the solar radiation at ambient temperature condition.

The total output power obtained from the SPV arrays can be mathematically expressed as:

$$P_{SPV}^{max}(t) = N_{SPV} \times \mathrm{P_{SPV}(t)} \quad (2)$$

2) MATHEMATICAL MODELING OF WIND TURBINE

A wind turbine's power output is determined by wind speed and hub height. These factors can be computed using the following equation [48]:

$$\frac{V_2}{V_1} = \left(\frac{H_2}{H_1}\right)^{\mu_{WT}} \quad (3)$$

Where V_1 denotes the wind speed at height H_1 and V_2 represents the wind speed at hub height H_2, μ_{WT} is the coefficient of friction and its value is taken to be 0.143 in this work.

WT power is not directly proportional to wind speed. The output power of each WT unit is determined by several elements, including the turbine's rated power, cut-in wind speed, cut-out wind speed, and rated wind speed.

$$\mathrm{P_{WT}} = \begin{cases} 0 & \mathrm{V(\tau) \le V_{cut-in} \text{ or } V(\tau) \ge V_{cut-off}} \\ \mathrm{P_{WT-rated}} \frac{\mathrm{V(\tau) - V_{cut-in}}}{\mathrm{V_{rated} - V_{cut-in}}} & \mathrm{V_{cut-in} < V(\tau) < V_{rated}} \\ \mathrm{P_{WT-rated}} & \mathrm{V_{rated} < V(\tau) < V_{cut-off}} \end{cases} \quad (4)$$

In the given context, V (τ) represents the wind speed at a specific height, while $\mathrm{V_{cut-in}}$ and $\mathrm{V_{cut-off}}$ represent the cut-in and cut-off speeds, respectively. $\mathrm{P_{WT-rated}}$ refers to the rated power of the WT, and $\mathrm{V_{rated}}$ represents the rated wind speed.

3) MATHEMATICAL MODELING OF RENEWABLE OUTPUT POWER

Efficiency parameters in a model represent the energy losses during the transformation process. When calculating the total renewable energy produced by WT and SPV units, the efficiencies of AC-to-DC and DC-to-DC converters are taken into account.

$$P_{RE}(t) = \eta_{dc\ to\ dc} \times \sum_{n=1}^{N_{SPV}} P_{SPV}(t) + \eta_{ac\ to\ dc} \times \sum_{n=1}^{N_{WT}} P_{WT}(t) \quad (5)$$

In this research work, the efficiency of DC to DC and AC to DC conversion is considered to be 95% [49].

As a result of solar radiation and wind speed unpredictability, the energy output of SPV and WT systems can be inconsistent. SPV units must be integrated with BES systems in order to regulate the power supply, especially when charging electric vehicles. BES systems contribute to enhanced energy supply reliability through the storage of

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

surplus energy during periods of high generation and subsequent release in accordance with charging and discharging patterns.

4) MODELING OF DIESEL GENERATOR (DG)

Introducing DGs into the energy system can enhance its reliability and efficiency. DGs serve as backup sources of energy, particularly useful during peak demand or in cases of battery degradation. However, it's crucial to avoid running the DGs when they are unloaded or lightly loaded to ensure optimal efficiency and longevity.

When devising a hybrid energy system, it is crucial to consider the efficacy and hourly fuel consumption of the DG. These terms can be defined using the subsequent formula [50]:

$$u(t) = a * P_{DG}(t) + b * P_{rated-DG} \quad (6)$$

Here, $u(t)$ is the fuel consumption in liters per hour, $P(t)$ is the produced power in kW, $P_{rated-DG}$ is the rated power, and a and b are constant parameters in litre per kW that characterize the coefficients of fuel consumption and can be approximated to 0.246 and 0.08415, respectively.

The efficiency of DG can be computed by

$$\eta_{overall} = \eta_{brake\ thermal} * \eta_{DG} \quad (7)$$

Here, $\eta_{overall}$ is the overall efficiency of DG whereas $\eta_{brake\ thermal}$ stands for the DG's brake thermal efficiency.

5) MODELING OF BATTERY ENERGY STORAGE

This research examines the application of a BES in resolving demand-side power imbalances, thereby increasing the dependability and consistency of the electricity supply. Described below is the operation of the BES within the hybrid energy system.

When the power generated by the different components of the system surpasses the demand of the load, the surplus power is directed to the BES, where it initiates the charging procedure. The charging behavior of the BES at a specific moment (τ) is mathematically represented as follows: The BES receives the surplus power allocated to it when the power produced by the different components of the system surpasses the demand of the load. This causes the charging cycle to commence. The subsequent expression represents the charging behavior of the BES at a specific instant τ mathematically [51]:

$$SOC_{bat}(\tau+1) = SOC_{bat}(\tau)(1-\tilde{\lambda}) + \left[\left(P_{SPV}(\tau) + \frac{P_{WT}(\tau)}{\eta_{bdinv}}\right) - P_{EVCS}(\tau)\right] * \eta_{bat} \quad (8)$$

In situations where the RES fail to provide the EVs with the necessary power, the BES is triggered to supply the necessary power. The BES initiates discharge in accordance with the demand. Equation (9), which characterizes the discharge behavior of the BES at a specific moment τ, can be applied.

$$SOC_{bat}(\tau+1) = SOC_{bat}(\tau)(1-\tilde{\lambda}) - \frac{\left[P_{EVCS}(\tau) - \left(P_{SPV}(\tau) + \frac{P_{WT}(\tau)}{\eta_{bdinv}}\right)\right]}{\eta_{bat}} \quad (9)$$

In Eqn. (9). $SOC_{bat}(\tau+1)$represents the BES's state of charge (SOC) at a specific instant (τ+1), while $SOC_{bat}(\tau)$ represents the SOC of the BES at the same instant. The efficiency of the bi-directional inverter is denoted by η_{bdinv}, the output power of the SPV modules is denoted by $P_{SPV}(\tau)$, the output power of the WT unit is denoted by $P_{WT}(\tau)$, and the battery's efficiency is reflected in η_{bat}.

6) MODELING OF BI-DIRECTIONAL INVERTER

The bidirectional converter plays a crucial role in the efficient transfer of power in the system, allowing for the conversion of DC power from sources like batteries or renewable energy systems into AC power for consumption, and vice versa for excess AC power. When sizing an inverter, two power measurement parameters are considered: continuous power, which denotes the inverter's ability to handle a continuous load, and rated power, which represents the inverter's capacity to withstand short-term power surges. The effectiveness of a bidirectional inverter can be mathematically expressed as follows [52]:

$$\eta_{bdinv} = \frac{P_i}{P_i + P_{int} + fP_i^2} \quad (10)$$

Where, P_i, P_{int} and f can be given as follows:

$$P_i = \frac{P_{out}}{P_n} \quad (11)$$

$$P_{int} = 1 - 99\left(\frac{10}{\eta_{10}} - \frac{1}{\eta_{100}} - 9\right)^2 \quad (12)$$

$$f = \frac{1}{\eta_{100}} - P_{int} \quad (13)$$

Where, η_{10} and η_{100} represents the efficiency 10% and 100% of nominal power.

7) MODELING OF EV LOAD

The load requirement of EV can be formulated as follows:

$$PD_{EV}^{k} = PD_{EV,dch}^{k} \times \eta_{dch}^{EV} \quad \forall\, k = [k_a\ k_d] \quad (14)$$

The constraints related to EV charging and discharging for the time interval of arrival time (k_a) and departure time (k_d) are given by Eqn. (15) and (16) respectively

$$PD_{EV,ch}^{k} \le EV_{ch}^{k} \times bat_{cap}^{EV} \times \zeta \quad \forall\, k = [k_a\ k_d] \quad (15)$$

$$PD_{EV,dch}^{k} \le EV_{dch}^{k} \times bat_{cap}^{EV} \times (1-\zeta) \quad \forall\, k = [k_a\ k_d] \quad (16)$$

The maximum value of SOC of EV can be computed using Eqn. (17).

$$SOC_{EV}^{k} = SOC_{EV}^{k-1} + \left\{\eta_{ch}^{EV} EV_{ch}^{k} - EV_{dch}^{k} \times \frac{1}{\eta_{dch}^{EV}}\right\} \times \Delta k \quad \forall\, k = [k_a\ k_d] \quad (17)$$

η_{ch}^{EV} and η_{dch}^{EV} represents the charging and discharging efficiency of the EV battery. The depth of discharge of EV battery affects the charging and discharging efficiency.

The time required for charging the EV (T_{ch}^{EV}) is dependent on arrival time, EV battery type and SOC of each EV (SOC_{EV}) and can be calculated using Eqn. (18).

$$T_{ch}^{EV} = \frac{1 - SOC_{EV}}{P_{ch}^{m}} \times bat_{cap}^{EV} \quad (18)$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

The battery capacity of each EV can be represented by $\mathrm{bat}_{\mathrm{cap}}^{\mathrm{EV}}$ and power of m^{th} charging point is denoted by P_{ch}^{m}.

The daily total number of EVs has also been used to calculate the total EV demand. Consequently, the EV demand is computed using Eqn. (19) as follows:

$$\mathrm{PD}_{\mathrm{EV}}^{\mathrm{k}} = \sum_{m=1}^{N_{PEV}} P_{ch}^{m} \tag{19}$$

8) MODELING OF UTILITY GRID

When demand in the first phase outpaces the supply of power from various energy sources and the capacity of batteries for storage, the excess energy will be borrowed from the utility grid and can be stated as follows [53]:

$$\begin{aligned} P_P^{grid}(t) = P_D^{EV}(t) \\ - \Big[P_{SPV-out}^{total}(t) + P_{DG}(t) \\ + \Big[\Big(P_{SPV-out}^{total}(t) \\ + \big(SOC_{bat}(t) - SOC_{bat}^{min}(t) \big) \Big) \\ \times \eta_{bdinv} \Big] \Big] \end{aligned} \tag{20}$$

where $\llbracket SOC_{bat}^{min}(t)$ indicates the battery's minimal state of charge and $P_P^{grid}(t)$ is the amount of energy deficiency that must be borrowed from the grid in kWh.

The excess energy will be sold to the utility grid in the second phase, which may be described as follows, when demand is less than the amount of power generated from the various energy sources and the battery is fully charged:

$$\begin{aligned} P_S^{grid}(t) = \Big[P_{SPV-out}^{total}(t) + P_{DG}(t) \\ + \Big[\Big(P_{SPV-out}^{total}(t) \\ - \big(SOC_{bat}^{max}(t) - SOC_{bat}(t) \big) \Big) \\ \times \eta_{bdinv} \Big] \Big] - P_D^{EV}(t) \end{aligned} \tag{21}$$

where $SOC_{bat}^{max}(t)$ indicates the battery's maximum state of charge and $P_S^{grid}(t)$ is the additional energy to be sold to the grid in kWh.

III. OPTIMIZATION FRAMEWORK

The primary objective of this research is to enhance the SPV/WT/Grid system in order to reduce the TNPC and LCOE, while concurrently guaranteeing sufficient reliability as measured by the Loss of Power Supply Probability (LPSP). This segment provides an extensive elucidation of the formulation process employed to establish operational constraints and objective functions for the optimization procedure. It also outlines the energy management strategy for achieving the optimal configuration. Furthermore, a novel application of metaheuristic-based optimization techniques is introduced to optimize the SPV/WT/Grid system.

A. TECHNICAL AND ECONOMIC OBJECTIVES

The main aim of optimizing the hybrid energy system is to decrease both the LCOE and the TNPC. This cost reduction objective includes improving the reliability of power supply, which is a key concern for energy system designers. The proposed method identifies the optimal amounts of photovoltaic panels, wind turbine capacity, and battery storage. MATLAB is employed to simulate the objective function, TNPC, and LCOE, while considering the relevant constraints and limitations.

In order to ascertain the TNPC, the net present cost of replacement costs, operation and maintenance expenses, and capital expenditures are taken into account. It can be mathematically written as follows:

$$TNPC = NPC_{cap} + NPC_{O\ and\ M} + NPC_{rep} \tag{22}$$

$$NPC_{cap} = N_{SPV} \times C_{SPV}^{cap} + N_{WT} \times C_{WT}^{cap} + N_{BES} \times C_{BES}^{cap} + N_{bdinv} \times C_{bdinv}^{cap} \tag{23}$$

$$NPC_{O\ and\ M} = PWA \times \left(N_{SPV} \times C_{SPV}^{O\ and\ M} + N_{WT} \times C_{WT}^{O\ and\ M} + N_{BES} \times C_{BES}^{O\ and\ M} + N_{bdinv} \times C_{bdinv}^{O\ and\ M} \right) \tag{24}$$

$$PWA = \frac{(1+IR)^N - 1}{IR \times (1+IR)^N} \tag{25}$$

$$NPC_{rep} = \lambda \times \left(N_{SPV} \times C_{SPV}^{rep} + N_{WT} \times C_{WT}^{rep} + N_{BES} \times C_{BES}^{rep} + N_{bdinv} \times C_{bdinv}^{rep} \right) \tag{26}$$

$$\lambda = \sum_{N=1}^{z} \frac{1}{(1+IR)^{N \times L}} \tag{27}$$

Where, C_{SPV}^{cap} , C_{WT}^{cap} , $C_{BES}^{cap}$and C_{biinv}^{cap} denotes the cost of each SPV, WT, BES and inverter unit, $C_{SPV}^{O\ and\ M}$, $C_{WT}^{O\ and\ M}$ and $C_{BES}^{O\ and\ M}$ represents the operation and maintenance cost of each SPV, WT and BES unit, C_{SPV}^{rep}, C_{WT}^{rep}, C_{BES}^{rep} and C_{bdinv}^{rep} are the replacement cost of each SPV, WT, BES and inverter unit. PWA and λ are factors used to transform operational and replacement costs to a single present cost. IR represents the interest rate, N represents the project's lifespan, and L represents the number of component replacements over the project's tenure.

The LCOE serves as a valuable metric for assessing the economic viability of an interconnected energy system. It is computed by dividing the annualized cost of system components by the total energy production required to meet capacity needs, as illustrated in Eq. (28).

$$LCOE = \frac{TNPC \times PWA}{\sum_{\tau=1}^{T} P_{Gen}(\tau)} \tag{28}$$

LCOE is influenced by a multitude of factors, such as initial investment, solar irradiation, tenure, operation, and maintenance expenses, PWA, degradation of the utilized SPV units, and more.

B. OPERATIONAL CONSTRAINTS

1) RELIABILITY CONSTRAINT

In this research, the LPSP is considered a vital constraint that must be met during the design phase of the hybrid renewable energy system to ensure uninterrupted energy supply. LPSP is quantified on a scale from 0 to 1, where 0 signifies that the generated power fully meets the load demand, and 1 indicates that the load demand cannot be

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

fulfilled. Thus, LPSP serves as a metric for assessing the reliability of the hybrid system, as explained below:

$$\text{LPSP} = \frac{\sum_{\tau=1}^{T} LPS\ (t)}{\sum_{\tau=1}^{T} P_{load}(t).\Delta t} \quad (29)$$

$$PPS\ (t) = \frac{P_{load}(t)}{\eta_{\text{bdinv}}} - \left[P_{RES}.\Delta t + E_{BES}(t) - E_{BES}^{min}\right] \quad (30)$$

$$E_{BES}^{min} = (1 - DOD) \times E_{BES}^{max} \quad (31)$$

Hence, reliability plays a crucial role in ensuring a consistent and adequate supply, serving as an empirical indicator of optimized integrated systems. The storage system, in particular, significantly enhances reliability. During the optimal planning phase, the maximum LPSP reliability index is prioritized to establish an efficient hybrid electrical system.

$$\text{LPSP} \leq \text{LPSP}_{max} \quad (32)$$

It is assumed that the utmost value of LPSP is 1%.

2) DECISION VARIABLES RELATED CONSTRAINT

During the optimization process of an energy system, a balance is struck between the technical constraints related to reliability and the cost objectives. Decision variables must be defined to facilitate the implementation of the proposed optimization procedure. The constraints for the decision variables in the study, which define the upper and lower bounds of their quantities, are expressed by the equations below:

$$N_{SPV}^{min} \leq N_{SPV} \leq N_{SPV}^{max} \quad (33)$$

$$N_{WT}^{min} \leq N_{WT} \leq N_{WT}^{max} \quad (34)$$

$$E_{BES}^{min} \leq E_{BES} \leq E_{BES}^{max} \quad (35)$$

$$P_{biinv}^{min} \leq P_{biinv} \leq P_{biinv}^{max} \quad (36)$$

C. ENERGY MANAGEMENT APPROACH

The reliability of a power system is established when its capacity aligns with the demand requirement, as measured by the probability of power supply failure. The present study considers the electric vehicle (EV) power demand in the Riyadh region of Saudi Arabia. Energy requirements are met through the utilization of two main sources: SPV panels and WT. The following describes the discrepancy between the power produced by RES (SPV and WT) and the demand for EV loads:

$$\Delta P(t) = P_{SPV}(t) + P_{WT}(t) - P_{EV}^{load}(t) \quad (37)$$

When the power generated by RES is not sufficient for that of the demand, the shortfall is compensated by drawing power from the grid $P_{grid}^{borrow}(t)$. On the contrary, in the event that RES fail to meet the load demand, any remaining surplus power may be exported back to the grid denoted as $P_{grid}^{sold}(t)$. Nevertheless, purchasing and selling power to the grid are constrained, as indicated by the maximum purchase from the $P_{grid}^{borrow,max}(t)$ and the maximum sell to the $P_{grid}^{sold,max}(t)$.

- If $\Delta P(t) > 0$, the combined power output generated by the SPV panels and WT is adequate to charge the EVs, any surplus power $P_{grid}^{sold}(t)$ is determined to be sold to the grid. This surplus power available for sale to the grid is computed as follows:

$$(P_{grid}^{sold}(t)) = \left(P_{SPV}(t) + P_{WT}(t) - P_{EV}^{load}(t)\right) \quad (38)$$

- In a scenario where the output of a renewable energy source (RES) is sufficient to meet the power requirements of the electric vehicle (EV) load and exceeds the utmost selling capacity of the power grid, denoted as $\Delta P(t) > 0\ and\ \Delta P(t) > (P_{grid}^{sold,max}(t))$, the excess electrical energy is directed to the dump load. This surplus energy allocated to the dump load can be computed as follows:

$$P_{dump}(t) = P_{SPV}(t) + P_{WT}(t) - P_{EV}^{load}(t) - P_{grid}^{sold,max}(t)/\eta_{bdinv} \quad (39)$$

- If $\Delta P(t) > 0$, it indicates that the power generated by the SPV panels and WT is insufficient to meet the energy requirements of the EVs. Consequently, the required electrical energy is sourced from the grid network, as outlined below:

$$(P_{grid}^{borrow}(t)) = (P_{EV}^{load}(t) - P_{SPV}(t) - P_{WT}(t))/\eta_{bdinv} \quad (40)$$

- If $\Delta P(t) = 0$, it signifies that no power exchange occurs with the grid, and the power generated by the SPV panels and WT is adequate to satisfy the demand of the EV load.
- When the combined output from RES and grid power falls short of meeting the peak load demand, a power deficit occurs, which can be expressed as:

$$P_{def}(t) = P_{grid}^{borrow}(t) - P_{grid}^{borrow,max}(t) \quad (41)$$

To guarantee the dependable fulfillment of the EV load requirement, it is critical to reduce the power deficit $P_{def}(t)$ to zero while endeavoring to minimize the LPSP. Enforcing precise boundaries for LPSP is advantageous for resolving the optimization problem. One percent is the utmost LPSP limit for this investigation. Utilizing the flowchart depicted in Figure 6, the energy management strategy for grid-connected SPV and WT-based design of EV load can be clarified.

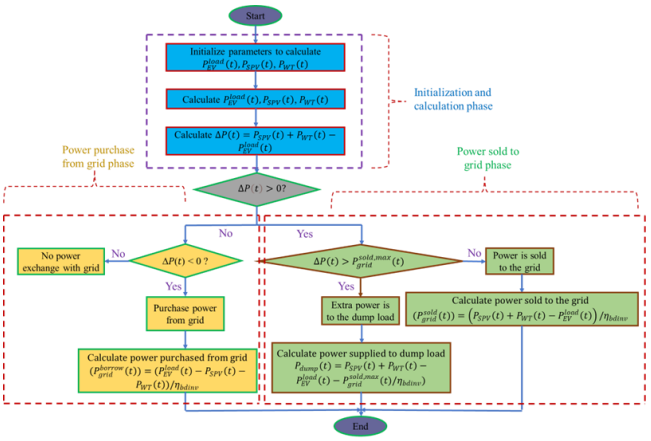


FIGURE 6. EMS for the optimal combination

C. OVERVIEW OF THE PROPOSED GIZA PYRAMID CONSTRUCTION ALGORITHM (GPCA)

In this study, the GPCA is employed to optimize the renewable energy system for fulfilling the EV demand in the Riyadh region of Saudi Arabia. The selection of this algorithm

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

is justified by its ease of implementation, efficient search capabilities, and the simplicity of its component algorithms.

1) OVERVIEW OF GPCA

The GPCA is a novel metaheuristic approach inspired by the ancient craftsmanship and architectural techniques used in the construction of the Giza Pyramids in Egypt [54]. This algorithm, based on population size, simulates the movements of laborers and the transportation of stone slabs up the slope under the supervision of a designated representative appointed by the Pharaoh. Demonstrated to handle multidimensional problems effectively, the GPCA is well-suited for optimizing both technical and economic objectives.

The core concept of the GPCA involves continuous pushing and movement of a stone block by workers to improve control. Unique positions and fitness values are assigned to a predetermined number of workers, determining their ideal locations. The movement and displacement of workers, akin to the shifting of stone blocks, generate new potential positions. Through a substitution probability, existing worker positions may be replaced with new ones. These positions are then sorted to identify the best solution. In the algorithm, the stone block's movement up the slope is governed by a random initial velocity u_0 ranging from 0 to 1. As the stone block advances, it encounters opposing kinetic friction force f_k, eventually coming to a stop after covering a certain distance d.

$$f_k = \mu_k mg\ cos\theta \tag{42}$$

where, m represents the mass of stone block, g is the acceleration due to gravity, $\mu_k$is the coefficient of friction, θ denotes the angle between slope and the horizontal surface.

Using the Newton second law of motion which can be represented as:

$$\vec{F} = m\vec{a} \tag{43}$$

Where, a denotes the acceleration.

Further, Eq. (43) can be expressed as follows:

$$ma = -mgsin\theta - f_k \tag{44}$$

The acceleration of stone block on the slope can be estimated using Eqn. (44) and Eqn. (45) as follows:

$$a = -g(sin\theta + \mu_k cos\theta) \tag{45}$$

The displacement covered by the stone block on the slope is represented by d and can be determined as follows:

$$d = \frac{{u_0}^2}{2g(sin\theta + \mu_k cos\theta)} \tag{46}$$

The position of worker to determine its position is calculated using Eqn. (47)

$$x = \frac{{u_0}^2}{2gsin\theta} \tag{47}$$

The updated position can be calculated using Eqn. (48)

$$\vec{P} = \left(\vec{P_l} + d\right) \times x\vec{\epsilon_l} \tag{48}$$

Where, $\vec{\epsilon_l}$ is an arbitrary vector which obeys levy distribution.

The probability of substituting workers can be given using Eqn. (49).

$$\rho_k = \begin{cases} \gamma_k & if\ rand[0,1] \le 0.5 \\ else\ \emptyset_k \end{cases} \tag{49}$$

Where $\emptyset$ and γ is the set of values of primary solutions and generated solutions respectively. The steps involved in the execution of GPCA is displayed in Fig. 7.

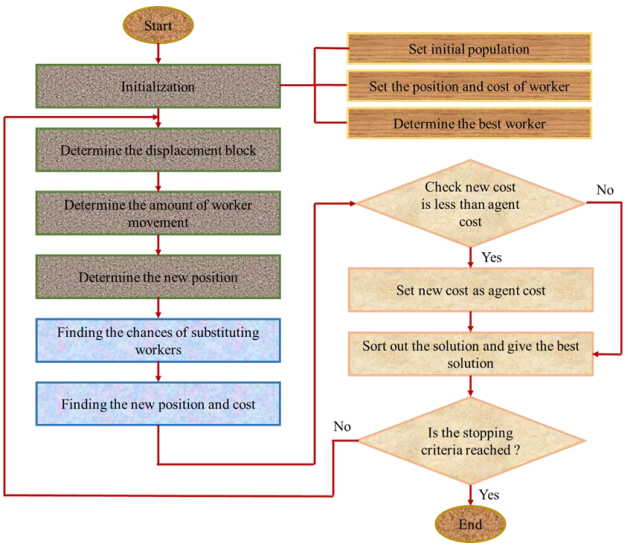


FIGURE 7. Flow chart of Giza Pyramids Construction Algorithm for the proposed combination

IV. RESULT AND DISCUSSIONS

This study endeavors to undertake a thorough technical and financial examination of renewable and grid-tied EVCS deployed in Riyadh, Saudi Arabia. It aims to capitalize on the plentiful RESs and the existing power network infrastructure in the area. Numerous optimization methodologies have been explored in previous literature to determine decision variables for fulfilling EV load demands while minimizing energy costs. In this research, the TNPC and LCOE of the proposed integrated EVCS framework are optimized utilizing the GPCA. The effectiveness and performance of GPCA are validated by comparing its outcomes with other optimization techniques such as the AHA, FPA, among others. Different system configurations, including combinations of energy resources and storage devices, are designed, investigated, and optimized to minimize electricity cost and TNPC. These configurations include SPV/WT/BES, SPV/DG/BES, WT/DG/BES, SPV/WT/Grid, and WT/DG/Grid. We examine the techno-economic variables for all feasible system configurations in Riyadh, Saudi Arabia. Overall, this research aims to provide insights into the optimal design and operation of EVCS integrated with renewable and grid-tied energy systems in Riyadh, KSA, considering various optimization approaches and system configurations.

In our study, we carefully considered the intermittent nature of RESs, which impacts their energy production and the bidirectional flow of electricity to and from the electric network. We simulated the energy system under various climatic seasons to account for these fluctuations. The design of the EVCS system model and the implementation of optimization techniques were conducted using the MATLAB environment. We analyzed several performance indices,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

including the LCOE, TNPC, system component capacities and productions, and grid sales and purchases for all possible system configurations. Additionally, we conducted sensitivity evaluations to assess the impact of varying solar irradiation, demand charge, wind speed, grid purchase price, discount, and inflation rates on TNPC and LCOE. Moreover, we investigated the impact of varying LPSP levels on both TNPC and LCOE. The convergence curve for the proposed GPCA and three other algorithms is illustrated in Fig. 8. We found that the computational time required by the GPCA technique was lower, with a small number of iterations needed to converge towards the optimal solution compared to the other approaches. Moreover, our analysis revealed that the GPCA technique consistently provided superior results for all levels of LPSP considered. Additionally, the operating cost of GPCA was the lowest compared to other techniques, as shown in Fig. 9, with a value of $2,845.42. Therefore, we propose the GPCA technique for further investigation and optimization of the energy-integrated EV charging station. The basic parameters used for tuning GPCA are displayed in Table 2, providing insights into the optimization process and parameter settings for future studies.

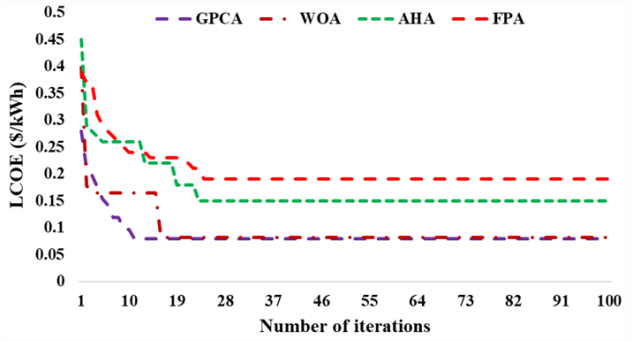


FIGURE 8. **Convergence curve of GPCA, WOA, AHA, and FPA for the selected site**

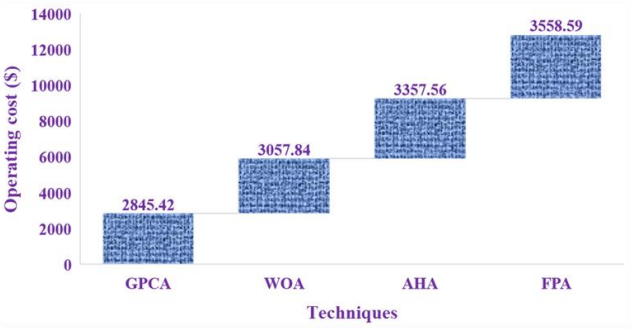


FIGURE 9. **Operating cost of various techniques**

TABLE 2. Basic parameters used for tuning of GPCA

METHOD	TUNED PARAMETERS
g	9.8
ρ_k	0.1
$f_{k,min}$	1
$f_{k,max}$	10

A. TECHNO-ENVIRON-ECONOMIC EVALUATION OF DIFFERENT SYSTEM CONFIGURATIONS FOR EVCS

1) PERFORMANCE ANALYSIS OF EVCS SYSTEM CONFIGURATIONS

In this subsection, we delve into the various system configurations for EVCS using performance indicator parameters. The electricity generated by RE systems exhibits significant fluctuations, leading to the incorporation of backup supply systems such as DGs, BES, and grid connectivity options to enhance reliability. Thus, the analysis considers possible system configurations designed for EV charging: SPV/WT/BES, SPV/DG/BES, WT/DG/BES, SPV/WT/Grid, and WT/DG/Grid. Table 3 presents the optimized results for all feasible system configurations aimed at meeting EV charging requirements in the specified area, achieved using the GPCA. A thorough examination of the table reveals that the SC#4 configuration (SPV/WT/Grid) exhibits the lowest TNPC of $102,042 and LCOE of $0.0796/kWh, making it the most desirable and cost-effective option compared to other configurations. The simulation outcomes suggest that the optimal setup, comprising SPV, WT, and grid integration (SPV/WT/Grid), effectively meets the demands of EV charging. As RESs exhibit intermittent energy production, the grid network acts as a backup to address any shortfall in energy supply, thereby augmenting reliability. The SPV and WT systems collectively contribute to 69.2% of the annual electricity production, with the remaining energy sourced from the electric power network. Moreover, the optimal configuration involves purchasing approximately 29,810 kWh of energy annually from the grid while exporting around 30,124 kWh/year of electricity to the grid through sales. This highlights the competitive renewable penetration of 67.9% achieved by SC#4, emphasizing its efficacy in leveraging RES for EV charging infrastructure.

Furthermore, the analysis delves into the detailed examination of system configurations for EVCS using performance indicator parameters. The electricity generated by renewable energy systems exhibits significant fluctuations. Hence, various backup supply systems such as DGs, BES, and grid connectivity options are often employed to enhance the reliability of such systems. In this analysis, the potential system configurations designed for EV charging are SPV/WT/BES, SPV/DG/BES, WT/DG/BES, SPV/WT/Grid, and WT/DG/Grid. Table 3 showcases the optimized results for all feasible system configurations to meet the EVs' charging requirements in the considered area. These results, obtained using the GPCA, highlight that the SC#4 (SPV/WT/Grid) configuration demonstrates the lowest TNPC of $102,042 and LCOE of $0.0796/kWh, making it the most desirable and cost-effective option compared to other feasible configurations. Additionally, the optimal configuration encompasses more energy sources compared to SC#1, SC#2, and SC#3 models, signifying its reliability in electrification. While SC#1 (SPV/WT/BES) boasts a 100% renewable fraction, indicating its preference from an environmental standpoint, it records the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

highest TNPC ($985,235.92) and LCOE ($1.02/kWh), making it financially unviable for EV charging. Conversely, configurations interconnected with the grid, like SC#4, sell more electricity to the grid than they purchase, unlike SC#5 (WT/DG/Grid), which primarily imports its charge demand from the grid. This suggests that the optimal system can generate additional revenues for the EVCS operator. Furthermore, SC#2 integrates SPV, DG, and a storage battery, with solar PV contributing to the majority of energy production and achieving a renewable penetration of 41.5%. However, its high LCOE of $0.509/kWh and TNPC of $424,028.4 render it economically unfeasible for EVCS. Substituting SPV with WT in SC#2 results in SC#3, with reduced LCOE and TNPC of $0.4231/kWh and $362,877, respectively. SC#3 exhibits the highest environmental impact due to DG usage, alongside the least renewable penetration. Comparatively, the optimized system (SC#4) showcases the lowest emissions profile after SC#1, which has no carbon footprint. It significantly reduces emissions for all pollutant emission parameters compared to SC#2, SC#3, and SC#5 systems. SC#3 records the highest total emissions rate due to high fuel consumption of the DG, while SC#2 and SC#5 emit considerable quantities of carbon dioxide emissions.

TABLE 3. Optimized comparative results of system configurations for EVCS

Configurations for EVCS model	SPV/WT/BES	SPV/DG/BES	WT/DG/BES	SPV/WT/Grid	WT/DG/Grid
EVCS System Scenarios	SC#1	SC#2	SC#3	SC#4	SC#5
LCOE ($/kWh)	1.02	0.5090	0.4231	0.0796	0.2110
TNPC ($)	985235.92	424028.4	362877	102042	439909
Operating cost ($)	26980.48	24661.73	21399.80	2845.42	8367.71
Number of SPV used	78	36	–	27	–
Number of WT used	24	–	9	9	13
Battery units	242	232	216	–	–
Converter capacity (kW)	10.4	35.1	10.8	13	13.7
Renewable fraction (%)	100	41.5	21.2	67.9	21.6
SPV production (kWh/year)	137311 (73.3 %)	62386 (63 %)	–	48037 (49.7 %)	–
Wind production (kWh/year)	50135 (26.7 %)	–	18801 (27.6 %)	18826 (19.5 %)	17338 (25.4 %)
DG production (kWh/year)	–	36701 (37 %)	49440 (72.4 %)	–	–
Grid purchase (kWh/year)	–	–	–	29810 (30.8 %)	50845 (74.6 %)
Grid sales (kWh/year)	–	–	–	30124 (32.4 %)	2109 (3.25 %)
CO2 emissions (kg/year)	0	37893	47071	18840	32134

TABLE 4. Environmental emissions for EVCS system configurations

Pollutants	SPV/WT/BES	SPV/DG/BES	WT/DG/BES	SPV/WT/Grid	WT/DG/Grid
Carbon Dioxide (kg/year)	0	37893	47071	18840	32134
Carbon Monoxide (kg/year)	0	239	356	0	0
Unburned Hydrocarbons (kg/year)	0	10.4	13	0	0
Particulate Matter (kg/year)	0	1.45	21.6	0	0
Sulphur Dioxide (kg/year)	0	92.8	115	81.7	139
Nitrogen Oxides (kg/year)	0	224	405	39.9	68.1

2) COST-WISE BREAKUP OF DIFFERENT EVCS SYSTEMS

A detailed breakdown of the costs associated with different components for the system configurations designed to meet the EVs' charging requirements in the selected location is provided in Fig. 10. In the case of SC#1, the majority of expenses are attributed to the BES and the WT. Notably, the capital cost of the BES ($266,200) exceeds double that of the WT ($106,800), resulting in higher replacement costs for these components compared to the SPV system and converter. The expenditure on purchasing the WT is 6.25% higher than that of the SPV system. Analyzing the cost breakdown of the components for the SPV/WT/BES system in detail, the capital cost ($481,824.2) and the replacement cost ($339,623) stand out as the most significant cost parameters. These costs are notably higher compared to other cost variables. The substantial capital cost can be attributed to the significant expenditure on the BES and WT, with the BES alone

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

contributing more than half of the total capital cost, while the converter has the smallest share ($8,306) of this cost. Specifically, the BES accounts for approximately 55.2% of the system's gross capital cost, while the WT has approximately 22.2%. Overall, the high capital cost observed in the SPV/WT/BES system is primarily due to the substantial expenditure on the BES and WT, highlighting their importance in the system.

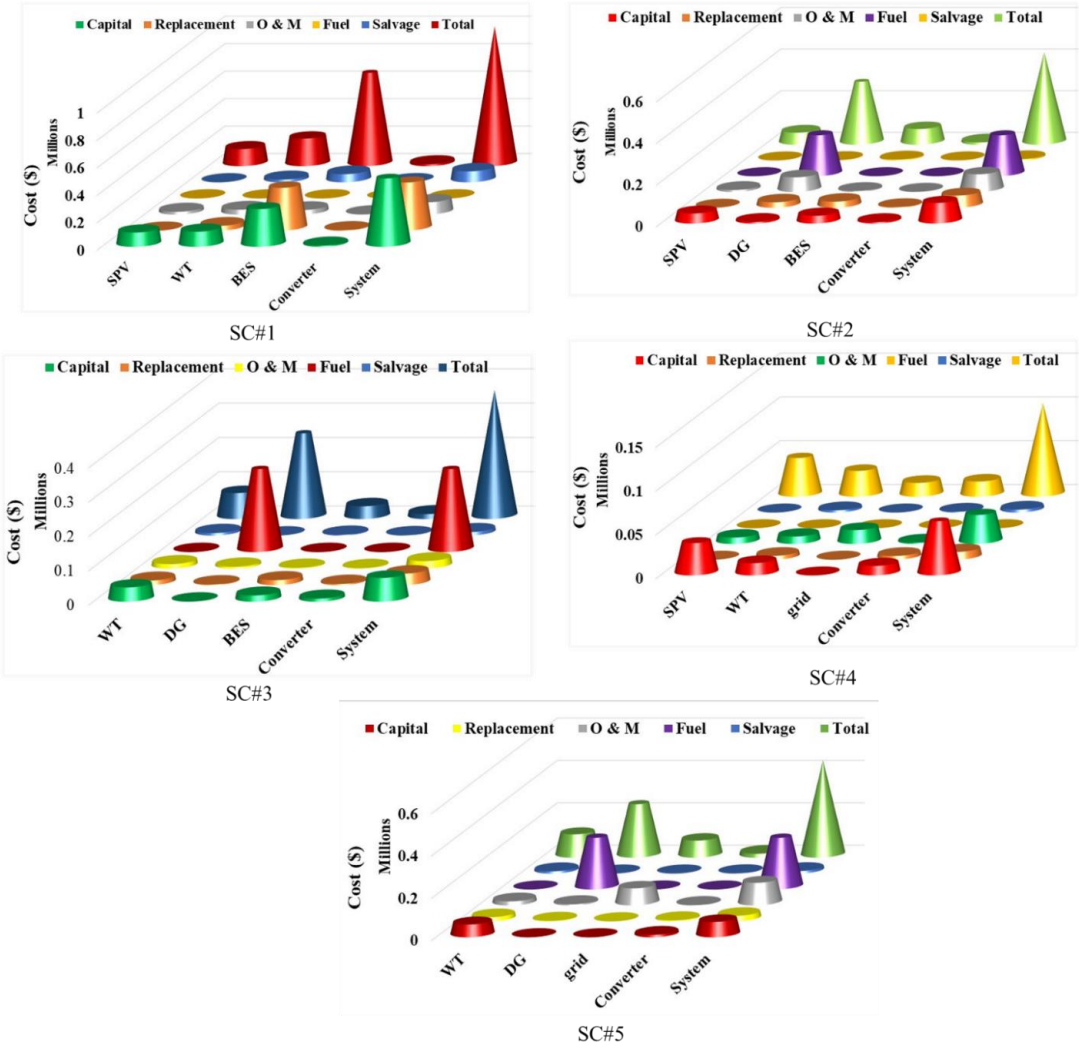


FIGURE10. Cost-wise breakup of various components for the different possible system configurations

In contrast, the elevated system replacement cost primarily stemmed from the substantial expense incurred when replacing the battery storage device. Specifically, the replacement cost of the BES comprises approximately 89% of the system replacement cost, followed by the WT, which accounts for about 10%. The operational and maintenance (O&M) costs of the WT system and BES were slightly higher, totaling $86,477.14, with the converter contributing the least at $1,073. Unlike the WT and BES, the SPV system incurs zero replacement and salvage costs due to its lifespan, which aligns with the project's duration. It's noteworthy that the SPV/WT/BES system reported the highest TNPC of $985,235.92 among all feasible configurations for the EVCS due to the higher capital and replacement costs. These costs accounted for approximately 48.9% and 34.47% of the TNPC, respectively. In the case of SC#2, the cost breakdown reveals that fuel and capital costs account for approximately 44.1% and 22.2% of the TNPC, respectively. The high fuel cost is attributed to expenses associated with DG fuel resources and operating conditions. Additionally, the high capital cost is due to substantial expenditures on SPV and BES devices. The DG's maintenance and operating conditions significantly contribute to high O&M costs, approximately 19.2% of the TNPC for the configuration. Similarly, the replacement cost accounts for approximately 13.2% of the TNPC, largely contributed by the BES and DG, with no replacement costs incurred by the SPV. Consequently, the TNPC of SC#2 is $424,028.4, lower than that of SC#1 and SC#5.

The total system capital cost (TSCC) for the SC#3 configuration is reported at $66,645. Within this cost

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

breakdown, the WT capital cost comprises approximately 60.1% of the TSCC, followed by the BES at about 26.4%. Conversely, the DG represents the least expensive component, accounting for only 0.56% of the TSCC. Regarding the overall replacement cost of $32,311, the BES constitutes around 43.7%, the WT about 39.5%, and the converter approximately 10.6% of this cost. The DG contributed the lowest cost at $1,978, while the storage device accounted for the highest share of the total replacement cost due to its shorter operational lifespan compared to the system's overall duration. Maintenance and operating conditions of the WT resulted in approximately 60.4% of the overall O&M cost. The total fuel (64.2%) and capital (18.4%) costs together constitute about 82.5% of the TNPC. This is primarily due to the significant expenditure on the WT and the storage device, as well as the expenses associated with the fuel resources for the DG.

For the SC#4 optimal configuration, the capital cost represents the highest share of the TNPC, with SPV and WT constituting approximately 59.7% and 22.7% of the overall capital cost, respectively. In terms of operating and maintenance (O&M) costs, the grid network dominates, accounting for approximately 47.8% of the total O&M cost, while the converter contributes the least share at around 4.25%. Due to the absence of DG in this system configuration, no expenditure is recorded on fuel resources, resulting in a zero-fuel cost. Additionally, the zero-replacement cost reported by the SPV and the distribution grid components leads to a low system replacement cost of $8,390. As a result, the optimal system configuration reports the lowest TNPC of $102,042 compared to other feasible configurations. In the case of SC#5, the total fuel cost constitutes more than half (53%) of the TNPC, primarily due to expenses associated with fuel resources. This is followed by the overall O&M cost, which accounts for about 23%. The high total O&M cost is attributed to the operational condition and maintenance of the grid and the WT, with the grid alone representing approximately 75% of the overall O&M cost. Regarding the total capital and replacement costs for SC#5, in both instances, the WT makes the largest contribution, comprising 83.6% of the capital cost and 74.4% of the replacement cost. Conversely, the DG constitutes a negligible 8% of the total replacement cost and a mere 0.54% of the overall capital cost. Overall, the capital and replacement costs represent about 15.7% and 5.63% of the TNPC for this configuration, respectively.

The operating costs of various EVCS configurations are illustrated in Fig. 11. Analysis of the results indicates that the optimal configuration, SC#4, has the lowest operating cost of $2,845.42, significantly lower than that of SC#1 ($26,980.48), SC#2 ($24,661.73), and SC#3 ($21,399.8) configurations. Additionally, the operating cost of SC#5 ($8,367.71) is almost three times higher than that of the optimal system. Furthermore, replacing the WT in the SC#1 configuration with a DG to form the SC#2 system reduces the operating cost by approximately $2,319. Similarly, eliminating the SPV component in the SC#2 configuration and replacing it with a WT to create the SC#3 configuration further decreases the operating cost by around $3,262.

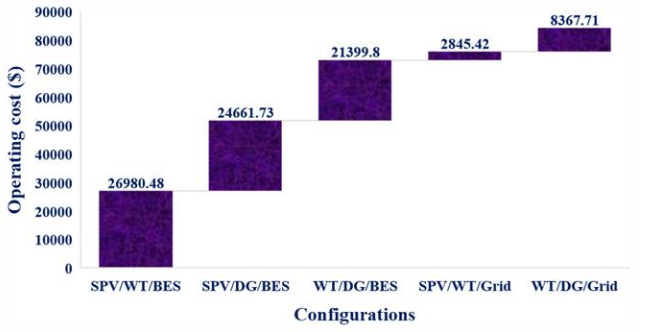


FIGURE 11. Operating cost of different EVCS system configurations

B. HOURLY SIMULATION RESULTS UNDER DIFFERENT CLIMATE SEASONS

The 24-hour profile of various energy components' power production, including grid purchase, for four different climatic seasons, is presented in Fig. 12. The seasons were categorized into four distinct cases: Case #1 represents the winter season, Case #2 the spring season, Case #3 the summer season, and Case #4 the autumn season. Upon closer examination of the figs, it is observed that there is generally lower power output from the SPV and WT systems from 12:00 AM to 6:00 AM for all seasons, except for the wind power output, which peaks during the summer season. The hourly simulation outcomes for the different climatic seasons are further explained below:

Case#1: During the winter hours, there is a decrease in output power from the WT with minimal solar output between 12:00 AM and 7:00 AM, which is not sufficient to meet the EV charging load. Consequently, additional energy is sourced from the grid to supplement the wind output and meet the EV charging demand adequately. During this period, the system experiences the highest energy purchase from the grid, resulting in zero revenue for the EVCS operator as no energy is sold back to the grid. Subsequently, after 7:00 AM, the output of solar and wind increases sufficiently, becoming the primary contributors to the charge requirement, with surplus energy sold back to the grid whenever generation exceeds EV load demand. Solar and wind outputs, along with energy sold to the grid, peak during the daytime (indicated by green, red, and yellow dashed lines in Fig. 12(a)). Consequently, the system purchases minimal electricity from the grid during this period (purple line in Fig. 12(a)). This trend continues until 6:00 PM, after which the system significantly increases its electricity purchases from the grid due to reduced power generation from SPV and WT systems.

Case#2: During the spring season, the system experiences minimal wind and solar power production in the early and late hours of the day, leading to a notable import of electricity from the grid to meet the EV load demand (illustrated by the light green dashed line in Fig. 12(b)). Similar to the winter season, the majority of solar and wind energy production occurs

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

during the middle of the day. During this period, solar and wind output adequately meets the EV charging demand, with surplus energy being injected into the grid. An examination of Fig. 12(b) reveals a significant surplus of exported energy, generating additional revenue for the charging system operator after satisfying the EV load. During this time, minimal electricity is withdrawn from the grid. The wind power output and grid sales are illustrated in Fig. 15(b) by the blue and green dashed lines, respectively.

Case#3: In the summer season, the optimal configuration exhibits a different pattern of solar and wind power production compared to the winter and spring seasons. As shown in Fig. 12(c), the WT reaches its maximum power output at the start of the day (indicated by the red dashed line), while the solar PV generates most of its power during the daytime (illustrated by the blue dashed line). Compared to previous seasons, there is minimal electricity drawn from the grid network during the summer period. However, to fulfill the load requirement, there are two distinct periods of heightened electricity import from the grid: between 5:00 PM and 11:00 PM, and between 3:00 AM and 7:00 AM. From the provided image, it's evident that the system generates excess energy after meeting the load demand during the day and early morning, attributed to the substantial renewable resource output. This surplus energy is subsequently sold to the electric network, capitalizing on advantageous incentives, thereby generating additional revenue for the charging system operator.

Case#4: The red dashed line depicted in Fig. 12(d) illustrates the WT commencement of energy production after 6:00 AM, maintaining relatively consistent power generation throughout the day. Similar to scenario #3, the bulk of the electricity drawn from the electric grid to supplement renewable energy output and fulfill the EV load occurs during the early and late hours of the day. Concurrently, the solar PV system predominantly generates power during daylight hours. As indicated by the purple dashed line in Fig. 12(d), during periods of surplus energy generation, the excess energy is sold to the power network operator once the system has met the EV charging requirement. This surplus energy sale contributes to additional income for the EVCS operator.

The comprehensive power analysis conducted for the optimal configuration of Riyadh's EV charging stations underscores the critical role of solar and wind energy sources in meeting the city's EV load demands. These RES effectively fulfilled the EV charging requirements to a satisfactory extent during periods of high solar and wind production. Nonetheless, a small amount of electricity was drawn from the grid to compensate for times when solar and wind energy production were minimal or absent. One notable aspect is the surplus electricity generated by the renewable components whenever their output exceeded the EV load demand. This surplus energy was effectively sold to the electric network at favorable incentives, providing an additional source of revenue for the EV charging station owners. This demonstrates the potential for renewable energy systems not only to meet local demand but also to contribute to the wider electric grid, fostering sustainability and economic viability in the energy sector.

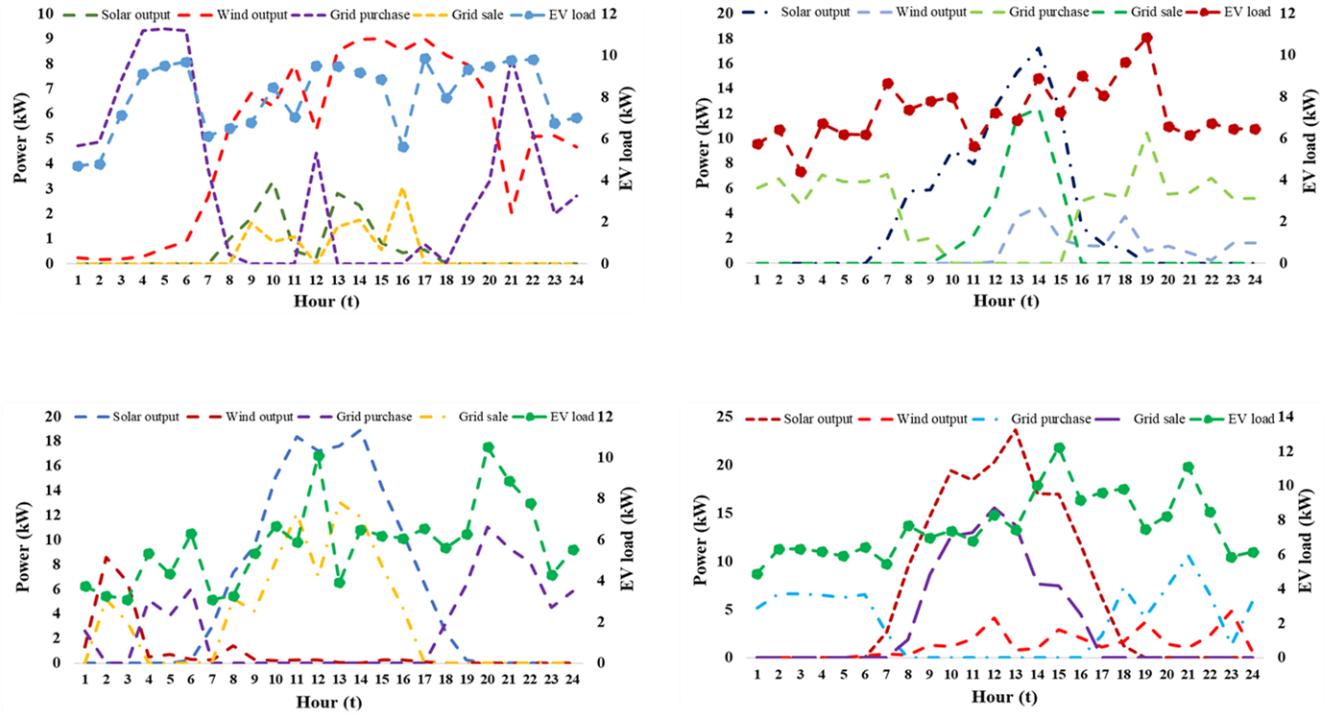


FIGURE 12. **Power assessment for SPV/WT/Grid configuration for EVCS during the (a) Winter (b) Spring (c) Summer and (d) Autumn season**

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

C. MONTHLY GRID POWER PURCHASE AND SALE

The monthly exchange of electricity with the distribution grid, as depicted in Fig. 13, showcases the system's ability to consistently generate surplus power throughout the year, particularly from March to October. This surplus energy was effectively sold to the electric network, generating additional revenue for the system operator. July witnessed the highest grid sales at 3041 kWh, while December recorded the lowest at 1937 kWh. Conversely, January and December saw the highest imports of electricity from the grid due to insufficient energy from renewable resources. However, in months like June and July, minimal power was imported from the grid as the SPV and WT systems provided ample electricity. In summary, these findings emphasize the technological and economic advantages of adopting a grid-tied renewable energy system. They highlight its capacity to utilize surplus energy production for financial benefits and ensure a dependable power supply, even in instances of low renewable energy generation.

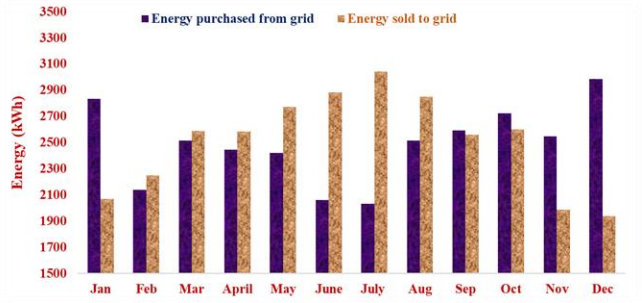


FIGURE 13. Monthly energy purchase and sale with the electric power network

D. IMPACT OF THE VARIATION IN LPSP ON TNPC AND LCOE

The comparative analysis presented in Table 5 highlights the impact of varying levels of LPSP on the TNPC and LCOE of EVCSs optimized using GPCA, WOA, AHA, and FPA techniques. Across all LPSP levels, the GPCA technique consistently yielded the lowest TNPC and LCOE compared to the other optimization methods. Specifically, at a 5% LPSP index, the GPCA approach achieved the lowest TNPC ($89,767.71) and LCOE ($0.053/kWh) among all techniques. The TNPC and LCOE of the GPCA-optimized results ranged between $89,767.71 to $95,692.75 and $0.053/kWh to $0.079/kWh, respectively, for varying LPSP levels. Notably, increasing the LPSP index percentage led to reductions in both TNPC and LCOE for the GPCA technique. Similar trends were observed with other optimization techniques, albeit with some variations. For instance, the LCOE of AHA and FPA increased by 16.4% and 22.2%, respectively, when the LPSP levels rose from 3% to 5%. In contrast, the GPCA system exhibited a reduction in LCOE by 15.2%, 8.95%, and 13.1% as the LPSP index increased from 0% to 1%, 1% to 3%, and 3% to 5%, respectively. Moreover, the TNPC value determined by the GPCA decreased by 2.21%, 1.08%, and 3.02% with increasing LPSP index levels from 0% to 1%, 1% to 3%, and 3% to 5%, respectively. These findings underscore the superior performance and robustness of the GPCA technique in optimizing EVCS configurations across varying levels of LPSP.

The comparison of optimized results at different LPSP levels highlights the superior performance of the GPCA technique across various scenarios. At a zero LPSP index, the SPV/WT/Grid configuration optimized by GPCA achieved a TNPC of $95,692.75, representing reductions of 7.6%, 17.2%, and 23.9% compared to the TNPC obtained using WOA, AHA, and FPA techniques, respectively. Similarly, the LCOE of $0.079/kWh for the GPCA-optimized system is notably lower, being 3.66%, 48.36%, and 60.1% less than that determined by WOA, AHA, and FPA techniques, respectively. Moving to a 1% LPSP, the GPCA-optimized system exhibited a further reduction in both TNPC and LCOE. The TNPC of $93,578.21 is 6.96%, 15.37%, and 23.1% lower than that of WOA, AHA, and FPA, while the LCOE of $0.067/kWh is 10.67%, 27.17%, and 39.1% less than the results obtained using WOA, AHA, and FPA approaches. At a 3% LPSP, the GPCA-optimized system continued to demonstrate its effectiveness, achieving a TNPC of $92,564.59 and an LCOE of $0.061/kWh. These values represent reductions of 6.27%, 13.29%, and 20.1% for TNPC, and 11.6%, 22.78%, and 38.38% for LCOE compared to the results obtained using WOA, AHA, and FPA techniques, respectively. Finally, at a 5% LPSP index, the GPCA-optimized system presented the lowest LCOE and TNPC values of $0.053/kWh and $89,767.71, respectively. These results highlight significant improvements over the outcomes obtained using WOA, AHA, and FPA techniques, with reductions ranging from 5.68% to 22.1% for TNPC and 12.48% to 56.2% for LCOE. Overall, these findings underscore the effectiveness and efficiency of the GPCA technique in optimizing EVCS configurations across varying levels of LPSP, outperforming other optimization methods in terms of both TNPC and LCOE.

E. IMPACT OF AVERAGE SOLAR RADIATION AND WIND SPEED

The study investigated the impact of varying solar irradiation and wind speed on the TNPC and LCOE of the system, recognizing the intermittent nature of RESs and its influence on energy production and associated costs. The findings, depicted in Fig. 14, highlight notable changes in TNPC and LCOE values across different levels of solar irradiation and wind speed. In the scenario of average solar irradiation (6.41 kWh/m2/day), both TNPC and LCOE exhibit reductions of approximately 17.86% and 13.3%, respectively, compared to the base case values (TNPC=$102,042 and LCOE=$0.0796/kWh). Further increasing solar irradiation to 6.91 kWh/m2/day results in additional reductions in TNPC and LCOE, with the LCOE decreasing by about 10% and the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

TABLE 5. Optimal results using GPCA, WOA, AHA, and FPA for different LPSP levels

LPSP	System Cost	Techniques GPCA	WOA	AHA	FPA
0 %	TNPC ($)	95692.75	103576.12	115583.56	125761.78
	LCOE ($/kWh)	0.079	0.082	0.153	0.198
1 %	TNPC ($)	93578.21	100583.56	110578.11	121765.43
	LCOE ($/kWh)	0.067	0.075	0.092	0.11
3 %	TNPC ($)	92564.59	98762.57	106754.12	115842.57
	LCOE ($/kWh)	0.061	0.069	0.079	0.099
5 %	TNPC ($)	89767.71	95178.21	102573.11	109548.59
	LCOE ($/kWh)	0.053	0.068	0.092	0.121

TNPC dropping by approximately 11% compared to the values observed at 6.41 kWh/m2/day. However, when the average solar radiation decreases to 4.91 and 5.41 kWh/m2/day, TNPC and LCOE values experience increases. Specifically, TNPC rises by approximately 12% and 2.1%, while LCOE increases by 21.86% and 8%, respectively, compared to the base case values. These findings underscore the sensitivity of TNPC and LCOE to variations in solar irradiation, emphasizing the importance of accurately assessing and accounting for such fluctuations in RESs when designing and optimizing energy systems, particularly those integrated with EV charging stations.

In the second scenario, depicted in Fig. 14(b), the graph illustrates the TNPC and LCOE as functions of varied wind speed. Similar to the first condition involving average solar irradiation, the TNPC and LCOE exhibit linear reductions as the average wind speed increases from 3.75 to 5.75 m/s. Specifically, there is a reduction of approximately 34.7% in TNPC and 36.56% in LCOE over this range of wind speeds. When the average wind speed is 5.25 meters per second, the TNPC and LCOE decrease by approximately 17.8 percent and 17.1 percent, in comparison to the values in the base case scenario (TNPC=$102,042 and LCOE=$0.0796/kWh). These reductions in TNPC and LCOE across varying wind speeds can be attributed to decreased fuel consumption and reduced costs associated with the electric grid network. The results emphasize the substantial influence of wind speed on the expenses of energy systems (TNPC and LCOE), thereby underscoring the criticality of incorporating and optimizing wind resources during the planning, construction, and operation phases.

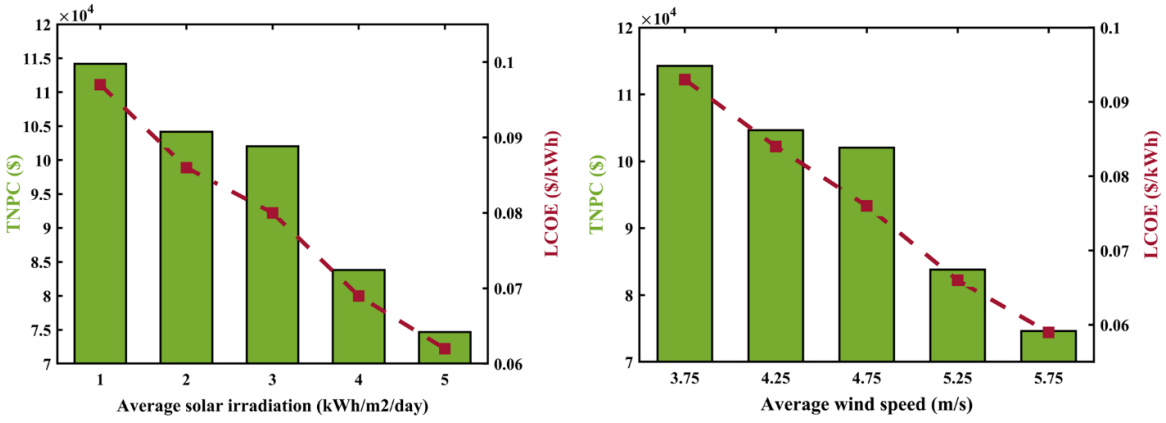


FIGURE 14. Effects of varying average (a) solar irradiation and (b) wind speed on TNPC and LCOE Monthly energy purchase and sale with the electric power network

F. IMPACT OF DISCOUNT AND INFLATION RATES

In this section, the impacts of varying discount rates from 9.09% to 13.09% and inflation rates from 6.5% to 10.5% on TNPC and LCOE are demonstrated in Fig. 15. It's crucial to assess the impact of varying these variables on the system cost, especially considering the uncertainty in the country's economy. The sensitivity analysis results depicted in Fig. 15(a) indicate that increasing the discount rate from 9.09% to 13.09% results in a decrease in TNPC and an increase in LCOE values. Conversely, the TNPC increases while the LCOE reduces as the inflation rate level rises, as depicted in Fig. 15(b). Specifically, the LCOE rises from $0.054/kWh for a 9.09% discount rate to $0.092/kWh as the discount rate increases to 13.09%. In contrast, the TNPC falls to $74,615.68 from $116,754.2 as the discount rate increases from 9.09% to 13.09%. This indicates a 70.3% increase in LCOE and a

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

36.1% decrease in TNPC due to the increment in the discount rate from 9.09% to 13.09%. Conversely, the TNPC increases from $74,654.45 at a 6.5% inflation rate to $115,754.2 when the inflation rate rises to 10.5%, while the LCOE drops from $0.095/kWh to $0.061/kWh for the same variation scenario. Notably, a rise in the inflation rate by approximately 12% from the baseline value (8.5%) is found to increase the TNPC by about 2.2% and reduce the LCOE by around 15.2%. These results underscore the sensitivity of TNPC and LCOE to variations in discount and inflation rates, highlighting the importance of considering economic factors in the planning and analysis of energy systems.

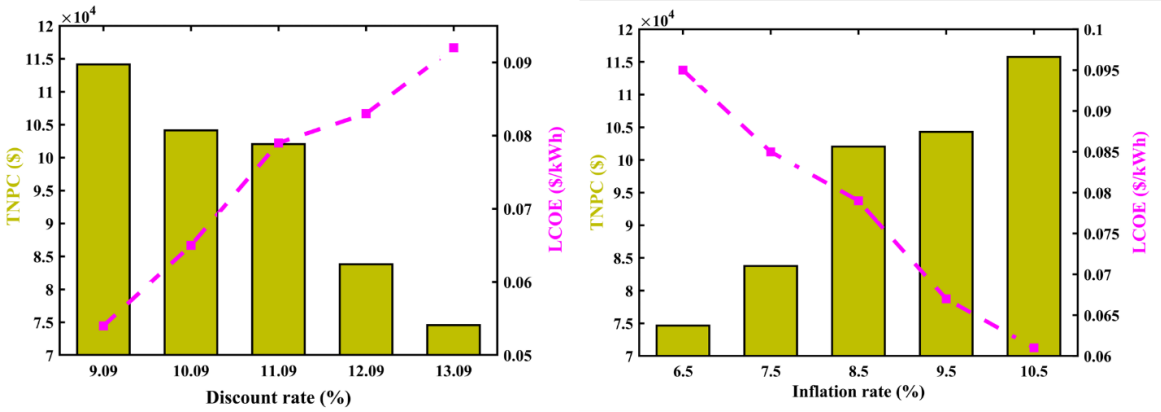


FIGURE 15. Impacts of varying (a) Discount rate and (b) Inflation rate on TNPC and LCOE

G. EFFECTS OF GRID PURCHASE PRICE AND DEMAND CHARGES

In this section, we delve into the impact of the electricity purchase price from national grid facilities and demand charges on both TNPC and LCOE. It has been demonstrated that TNPC and LCOE rise with increasing electric network purchase price and demand charges. As depicted in Fig. 16, an escalation in the cost of energy imported from the electric power network to meet the load requirement leads to an increase in both the LCOE and TNPC. For example, as the grid purchase price rises from $0.04 to $0.2 per kWh, the LCOE elevates from $0.054/kWh to $0.098/kWh, while the TNPC increases from $76,575.68 to $117,154.2. Similarly, an increase in the purchase price of about 33.3% from the base-case value ($0.12/kWh) causes the LCOE and TNPC to rise by approximately 10% and 5%, respectively. Conversely, reducing the baseline purchase price by 66.67% will reduce the optimal TNPC and LCOE by 25% and 31.6%, respectively. Additionally, the LCOE rises from $0.052/kWh when the demand charge is $0.7 per kW to $0.094/kWh as the demand charge increases to $2.7 per kW, while the TNPC increases from $74,575.68 to $114,154.2 for the same varying condition. The TNPC increases by 11.87%, and the LCOE rises by approximately 19% when the base-case demand charge increases by about 58.8%. These results emphasize the sensitivity of TNPC and LCOE to variations in the price of electricity purchase from the grid and demand charges, highlighting the importance of considering these factors in the economic analysis of energy systems.

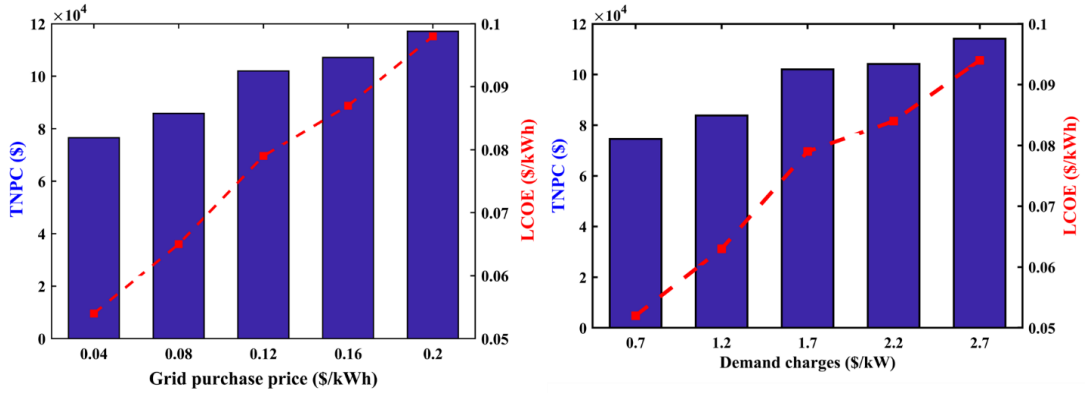


FIGURE 16. Effects of varying (a) grid purchase price and (b) demand charges on TNPC and LCOE.

V. CONCLUSION

Integrating renewable energy systems with the electrical vehicle charging station (EVCS) in Riyadh, Saudi Arabia, is a technologically and economically viable endeavor, as thoroughly examined in this study. Riyadh was selected due to its substantial supply of renewable energy sources, its crucial

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

contribution to the advancement of e-mobility, and its capacity to swiftly implement policies pertaining to electric vehicles. The objective of this research was to minimize the TNPC and LCOE while ensuring LPSP dependability through optimal component measurement using the GPCA. In contrast to alternative algorithms including the WOA, FPA, and AHA, this approach was utilized. The optimal system model, comprising solar PV, a WT, and the electric power network (SPV/WT/Grid), significantly reduced TNPC to $102,042 and LCOE to $0.0796/kWh. This configuration, labeled SC#4, achieved a remarkable 67.9% renewable penetration and effectively balanced grid energy purchases and sales. Conversely, the environmentally superior SC#1 (SPV/WT/BES) configuration with a 100% renewable fraction was economically unfeasible due to its high TNPC and LCOE. Configurations integrated with the grid, particularly SC#4, demonstrated more favorable energy transaction balances compared to others like SC#5 (WT/DG/Grid). The study also explored the impact of varying LPSP levels and diverse climatic, financial, and technical parameters on system costs, providing insights into the challenges, opportunities, and socio-economic aspects of EV implementation in KSA. The GPCA emerged as the most effective optimization technique, offering the lowest TNPC and LCOE values. Enhancements in wind speed or solar irradiance, along with reductions in demand charges or grid purchase prices, yielded significant financial advantages. Overall, the study provides valuable insights for policymakers and stakeholders involved in sustainable EV infrastructure development in Riyadh and beyond.

Despite the thoroughness of our study, we acknowledge certain limitations stemming from the scarcity of available data on EVs and charging infrastructure in KSA, as well as the early stage of EV adoption in the country. These limitations have led to certain assumptions in our analysis, which may affect the accuracy of our findings. To address these limitations, future research efforts should focus on collecting more comprehensive and accurate data on EV usage patterns, charging behavior, infrastructure requirements, and renewable energy potential in KSA. This would enable more precise modeling and optimization of EV charging systems, leading to more reliable results and actionable insights for policymakers and stakeholders. Despite these challenges, our study aims to contribute valuable insights to the development of policies that support the widespread adoption of EVs in Riyadh and across KSA. By aligning with the nation's Vision 2030 objectives of environmental sustainability and technological advancement, our findings can inform the strategic planning and implementation of EV charging infrastructure in the country. Furthermore, by identifying current challenges and proposing potential strategies, our study offers a framework for the optimal design and deployment of EV charging infrastructure. This framework can serve as a blueprint for policymakers and stakeholders in KSA and other developing countries seeking to expand their EV charging networks and accelerate the transition to sustainable transportation systems.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number 223202.

REFERENCES

[1] A. Shafiq et al., “Integration of solar based charging station in power distribution network and charging scheduling of EVs,” Front Energy Res, vol. 11, Jan. 2023, doi: 10.3389/fenrg.2023.1086793.

[2] A. AlHammadi, N. Al-Saif, A. S. Al-Sumaiti, M. Marzband, T. Alsumaiti, and E. Heydarian-Forushani, “Techno-Economic Analysis of Hybrid Renewable Energy Systems Designed for Electric Vehicle Charging: A Case Study from the United Arab Emirates,” Energies (Basel), vol. 15, no. 18, p. 6621, Sep. 2022, doi: 10.3390/en15186621.

[3] M. Khalid, F. Ahmad, and B. K. Panigrahi, “Design, simulation and analysis of a fast charging station for electric vehicles,” Energy Storage, vol. 3, no. 6, Dec. 2021, doi: 10.1002/est2.263.

[4] M. R. Khalid, M. S. Alam, M. Krishnamurthy, E. A. Al-Ammar, H. Alrajhi, and M. S. J. Asghar, “A Multiphase AC–DC Converter With Improved Power Quality for EV Charging Station,” IEEE Transactions on Transportation Electrification, vol. 8, no. 1, pp. 909–924, Mar. 2022, doi: 10.1109/TTE.2021.3120032.

[5] M. G. A. A. Ahmed, “Electric Vehicle Impact on the Sustainable Development Goals Considering Renewable Energy Sources Integration,” African Journal of Advanced Pure and Applied Sciences (AJAPAS), vol. 2, no. 2, pp. 227–234, 2023.

[6] M. Mustafa Kamal, I. Asharaf, and E. Fernandez, “Optimal renewable integrated rural energy planning for sustainable energy development,” Sustainable Energy Technologies and Assessments, vol. 53, p. 102581, Oct. 2022, doi: 10.1016/j.seta.2022.102581.

[7] A. Alsharif et al., “Impact of Electric Vehicle on Residential Power Distribution Considering Energy Management Strategy and Stochastic Monte Carlo Algorithm,” Energies (Basel), vol. 16, no. 3, p. 1358, Jan. 2023, doi: 10.3390/en16031358.

[8] J. Qi and L. Li, “Economic Operation Strategy of an EV Parking Lot with Vehicle-to-Grid and Renewable Energy Integration,” Energies (Basel), vol. 16, no. 4, p. 1793, Feb. 2023, doi: 10.3390/en16041793.

[9] M. Huda, T. Koji, and M. Aziz, “Techno Economic Analysis of Vehicle to Grid (V2G) Integration as Distributed Energy Resources in Indonesia Power System,” Energies (Basel), vol. 13, no. 5, p. 1162, Mar. 2020, doi: 10.3390/en13051162.

[10] H. Abubakr, A. Lashab, J. C. Vasquez, T. H. Mohamed, and J. M. Guerrero, “Novel V2G regulation scheme using Dual-PSS for PV islanded microgrid,” Appl Energy, vol. 340, p. 121012, Jun. 2023, doi: 10.1016/j.apenergy.2023.121012.

[11] R. Doosti, A. Rezazadeh, and M. Sedighizadeh, “Power and Energy Management Strategies for a Microgrid with the Presence of Electric Vehicles and CAES Considering the Uncertainty of Resources,” Processes, vol. 11, no. 4, p. 1156, Apr. 2023, doi: 10.3390/pr11041156.

[12] A. Asaad et al., “Multi-objective optimal planning of EV charging stations and renewable energy resources for smart microgrids,” Energy Sci Eng, vol. 11, no. 3, pp. 1202–1218, Mar. 2023, doi: 10.1002/ese3.1385.

[13] A. M. Elshurafa and N. Peerbocus, “Electric vehicle deployment and carbon emissions in Saudi Arabia: A power system perspective,” The Electricity Journal, vol. 33, no. 6, p. 106774, Jul. 2020, doi: 10.1016/j.tej.2020.106774.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

[14] T. Alquthami, A. Alsubaie, M. Alkhraijah, K. Alqahtani, S. Alshahrani, and M. Anwar, "Investigating the Impact of Electric Vehicles Demand on the Distribution Network," Energies (Basel), vol. 15, no. 3, p. 1180, Feb. 2022, doi: 10.3390/en15031180.

[15] F. Aljahdali and M. Abbod, "Design of smart energy generation and demand response system in Saudi Arabia," in 2017 52nd International Universities Power Engineering Conference (UPEC), IEEE, Aug. 2017, pp. 1–5. doi: 10.1109/UPEC.2017.8231965.

[16] M. M. Rahman, G. A. Ghazi, E. A. Al-Ammar, and W. Ko, "Techno-Economic Analysis of Hybrid PV/Wind/Fuel-Cell System for EVCS," in 2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), IEEE, Jun. 2021, pp. 1–6. doi: 10.1109/ICECCE52056.2021.9514084.

[17] O. S. Alzaid, B. Salim, J. Orfi, S. Khan, and H. Alshehri, "Hybrid Solar and Wind Power Generation in Saudi Arabia," Energy and Environment Research, vol. 10, no. 2, p. 25, Nov. 2020, doi: 10.5539/eer.v10n2p25.

[18] K. A. Khan, M. M. Quamar, F. H. Al-Qahtani, M. Asif, M. Alqahtani, and M. Khalid, "Smart grid infrastructure and renewable energy deployment: A conceptual review of Saudi Arabia," Energy Strategy Reviews, vol. 50, p. 101247, Nov. 2023, doi: 10.1016/j.esr.2023.101247.

[19] M. Bilal, F. Ahmad, and M. Rizwan, "Techno-economic assessment of grid and renewable powered electric vehicle charging stations in India using a modified metaheuristic technique," Energy Convers Manag, vol. 284, p. 116995, May 2023, doi: 10.1016/j.enconman.2023.116995.

[20] D. S. Abraham et al., "Fuzzy-Based Efficient Control of DC Microgrid Configuration for PV-Energized EV Charging Station," Energies (Basel), vol. 16, no. 6, p. 2753, Mar. 2023, doi: 10.3390/en16062753.

[21] F. Ahmad, I. Ashraf, A. Iqbal, I. Khan, and M. Marzband, "Optimal Location and Energy Management Strategy for EV Fast Charging Station With Integration of Renewable Energy Sources," in 2022 IEEE Silchar Subsection Conference (SILCON), IEEE, Nov. 2022, pp. 1–6. doi: 10.1109/SILCON55242.2022.10028897.

[22] W. S. Tounsi Fokui, L. Ngoo, and M. Saulo, "Optimal Integration of Electric Vehicle Charging Stations and Compensating Photovoltaic Systems in a Distribution Network Segregated into Communities," Journal of Advanced Engineering and Computation, vol. 6, no. 4, p. 260, Dec. 2022, doi: 10.55579/jaec.202264.380.

[23] R. Verma, S. K. Sharma, P. Singh, J. K. Bhutto, and A. R. A. Alharbi, "Analysis and Sizing of Charging Stations in Kota City," Sustainability, vol. 14, no. 18, p. 11759, Sep. 2022, doi: 10.3390/su141811759.

[24] M. Bilal and M. Rizwan, "Integration of electric vehicle charging stations and capacitors in distribution systems with vehicle-to-grid facility," Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, pp. 1–30, May 2021, doi: 10.1080/15567036.2021.1923870.

[25] M. Bilal, M. Rizwan, I. Alsaidan, and F. M. Almasoudi, "AI-Based Approach for Optimal Placement of EVCS and DG With Reliability Analysis," IEEE Access, vol. 9, pp. 154204–154224, 2021, doi: 10.1109/ACCESS.2021.3125135.

[26] A. Eid, O. Mohammed, and H. El-Kishky, "Efficient operation of battery energy storage systems, electric-vehicle charging stations and renewable energy sources linked to distribution systems," J Energy Storage, vol. 55, p. 105644, Nov. 2022, doi: 10.1016/j.est.2022.105644.

[27] S. M. B. Sadati, J. Moshtagh, M. Shafie-Khah, A. Rastgou, and J. P. S. Catalão, "Optimal Charge Scheduling of Electric Vehicles in Solar Energy Integrated Power Systems Considering the Uncertainties," in Electric Vehicles in Energy Systems, Cham: Springer International Publishing, 2020, pp. 73–128. doi: 10.1007/978-3-030-34448-1_4.

[28] F. Ahmad, A. Iqbal, I. Asharf, M. Marzband, and I. Khan, "Placement and Capacity of EV Charging Stations by Considering Uncertainties with Energy Management Strategies," IEEE Trans Ind Appl, pp. 1–10, 2023, doi: 10.1109/TIA.2023.3253817.

[29] S. E. Ahmadi, S. M. Kazemi-Razi, M. Marzband, A. Ikpehai, and A. Abusorrah, "Multi-objective stochastic techno-economic-environmental optimization of distribution networks with G2V and V2G systems," Electric Power Systems Research, vol. 218, p. 109195, May 2023, doi: 10.1016/j.epsr.2023.109195.

[30] O. Ekren, C. Hakan Canbaz, and Ç. B. Güvel, "Sizing of a solar-wind hybrid electric vehicle charging station by using HOMER software," J Clean Prod, vol. 279, p. 123615, Jan. 2021, doi: 10.1016/j.jclepro.2020.123615.

[31] I. Alsaidan, M. Bilal, M. Alaraj, M. Rizwan, and F. M. Almasoudi, "A Novel EA-Based Techno–Economic Analysis of Charging System for Electric Vehicles: A Case Study of Qassim Region, Saudi Arabia," Mathematics, vol. 11, no. 9, p. 2052, Apr. 2023, doi: 10.3390/math11092052.

[32] J. O. Oladigbolu, A. Mujeeb, A. A. Imam, and A. M. Rushdi, "Design, Technical and Economic Optimization of Renewable Energy-Based Electric Vehicle Charging Stations in Africa: The Case of Nigeria," Energies (Basel), vol. 16, no. 1, p. 397, Dec. 2022, doi: 10.3390/en16010397.

[33] J. O. Oladigbolu, A. Mujeeb, Y. A. Al-Turki, and A. M. Rushdi, "A Novel Doubly-Green Stand-Alone Electric Vehicle Charging Station in Saudi Arabia: An Overview and a Comprehensive Feasibility Study," IEEE Access, vol. 11, pp. 37283–37312, 2023, doi: 10.1109/ACCESS.2023.3266436.

[34] N. Ghosh, S. Mothilal Bhagavathy, and J. Thakur, "Accelerating electric vehicle adoption: techno-economic assessment to modify existing fuel stations with fast charging infrastructure," Clean Technol Environ Policy, vol. 24, no. 10, pp. 3033–3046, Dec. 2022, doi: 10.1007/s10098-022-02406-x.

[35] C. Marinescu, "Progress in the Development and Implementation of Residential EV Charging Stations Based on Renewable Energy Sources," Energies (Basel), vol. 16, no. 1, p. 179, Dec. 2022, doi: 10.3390/en16010179.

[36] A. Selim, M. Abdel-Akher, S. Kamel, F. Jurado, and S. A. Almohaimeed, "Electric Vehicles Charging Management for Real-Time Pricing Considering the Preferences of Individual Vehicles," Applied Sciences, vol. 11, no. 14, p. 6632, Jul. 2021, doi: 10.3390/app11146632.

[37] H. T. Nguyen and D.-H. Choi, "Distributionally Robust Model Predictive Control for Smart Electric Vehicle Charging Station With V2G/V2V Capability," IEEE Trans Smart Grid, vol. 14, no. 6, pp. 4621–4633, Nov. 2023, doi: 10.1109/TSG.2023.3263470.

[38] D. Yang, N. J. S. Sarma, M. F. Hyland, and R. Jayakrishnan, "Dynamic modeling and real-time management of a system of EV fast-charging stations," Transp Res Part C Emerg Technol, vol. 128, p. 103186, Jul. 2021, doi: 10.1016/j.trc.2021.103186.

[39] W. Khan, F. Ahmad, and M. S. Alam, "Fast EV charging station integration with grid ensuring optimal and quality power exchange," Engineering Science and Technology, an International Journal, vol. 22, no. 1, pp. 143–152, Feb. 2019, doi: 10.1016/j.jestch.2018.08.005.

[40] A. Almutairi, "Impact Assessment of Diverse EV Charging Infrastructures on Overall Service Reliability," Sustainability, vol. 14, no. 20, p. 13295, Oct. 2022, doi: 10.3390/su142013295.

[41] D. N. Araujo et al., "Optimum Design of On-Grid PV-BESS for Fast Electric Vehicle Charging Station in Brazil," in 2021 IEEE PES Innovative Smart Grid Technologies Conference - Latin America (ISGT Latin America), IEEE, Sep. 2021, pp. 1–5. doi: 10.1109/ISGTLatinAmerica52371.2021.9543071.

[42] B. Wang, P. Dehghanian, and D. Zhao, "Coordinated Planning of Electric Vehicle Charging Infrastructure and Renewables in Power Grids," IEEE Open Access Journal of Power and Energy, vol. 10, pp. 233–244, 2023, doi: 10.1109/OAJPE.2023.3245993.

[43] Z. Ullah et al., "Optimal scheduling and techno-economic analysis of electric vehicles by implementing solar-based grid-tied charging station," Energy, vol. 267, p. 126560, Mar. 2023, doi: 10.1016/j.energy.2022.126560.

[44] S. M. S., F. Titus, S. B. Thanikanti, S. S. M., S. Deb, and N. M. Kumar, "Charge Scheduling Optimization of Plug-In Electric Vehicle in a PV Powered Grid-Connected Charging Station Based

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411291

on Day-Ahead Solar Energy Forecasting in Australia," Sustainability, vol. 14, no. 6, p. 3498, Mar. 2022, doi: 10.3390/su14063498.

[45] N. Das, A. Haque, H. Zaman, S. Morsalin, and S. Islam, "Domestic Load Management With Coordinated Photovoltaics, Battery Storage and Electric Vehicle Operation," IEEE Access, vol. 11, pp. 12075–12087, 2023, doi: 10.1109/ACCESS.2023.3241244.

[46] J. Leijon and C. Boström, "Charging Electric Vehicles Today and in the Future," World Electric Vehicle Journal, vol. 13, no. 8, p. 139, Jul. 2022, doi: 10.3390/wevj13080139.

[47] M. Bilal, I. Alsaidan, M. Alaraj, F. M. Almasoudi, and M. Rizwan, "Techno-Economic and Environmental Analysis of Grid-Connected Electric Vehicle Charging Station Using AI-Based Algorithm," Mathematics, vol. 10, no. 6, p. 924, Mar. 2022, doi: 10.3390/math10060924.

[48] B. K. Das, R. Hassan, M. S. H. K. Tushar, F. Zaman, M. Hasan, and P. Das, "Techno-economic and environmental assessment of a hybrid renewable energy system using multi-objective genetic algorithm: A case study for remote Island in Bangladesh," Energy Convers Manag, vol. 230, p. 113823, Feb. 2021, doi: 10.1016/j.enconman.2020.113823.

[49] A. K. Karmaker, Md. R. Ahmed, Md. A. Hossain, and Md. M. Sikder, "Feasibility assessment & design of hybrid renewable energy based electric vehicle charging station in Bangladesh," Sustain Cities Soc, vol. 39, pp. 189–202, May 2018, doi: 10.1016/j.scs.2018.02.035.

[50] M. M. kamal, I. Ashraf, and E. Fernandez, "Sustainable electrification planning of rural microgrid using renewable resources and its environmental impact assessment," Environmental Science and Pollution Research, Apr. 2022, doi: 10.1007/s11356-022-19949-7.

[51] S. Mandal, B. K. Das, and N. Hoque, "Optimum sizing of a stand-alone hybrid energy system for rural electrification in Bangladesh," J Clean Prod, vol. 200, pp. 12–27, Nov. 2018, doi: 10.1016/j.jclepro.2018.07.257.

[52] B. K. Das, N. Hoque, S. Mandal, T. K. Pal, and M. A. Raihan, "A techno-economic feasibility of a stand-alone hybrid power generation for remote area application in Bangladesh," Energy, vol. 134, pp. 775–788, Sep. 2017, doi: 10.1016/j.energy.2017.06.024.

[53] P. Anand, M. Rizwan, and S. K. Bath, "Sizing of renewable energy based hybrid system for rural electrification using grey wolf optimisation approach," IET Energy Systems Integration, vol. 1, no. 3, pp. 158–172, Sep. 2019, doi: 10.1049/iet-esi.2018.0053.

[54] S. Harifi, J. Mohammadzadeh, M. Khalilian, and S. Ebrahimnejad, "Giza Pyramids Construction: an ancient-inspired metaheuristic algorithm for optimization," Evol Intell, vol. 14, no. 4, pp. 1743–1761, Dec. 2021, doi: 10.1007/s12065-020-00451-3.

FAYEZ ALANAZI is an assistant professor in the department of Civil Engineering at Jouf University. Dr Alanazi specializes in Traffic and Transportation Engineering, with focus on enhancing urban mobility and improving transportation systems. Dr Alanazi obtained his PhD in Civil Engineering from the University of Akron in Ohio, USA, in 2021. His Doctoral research focused on developing innovative solutions for traffic management and transportation planning. Prior to his PhD, he completed his Master of Science in Civil Engineering at the University of Dayton, also in Ohio, in 2016. With strong educational background in Civil Engineering, Dr. Alanazi earned his Bachelor od Science degree in Civil Engineering from Jouf University in Saudi Arabia in 2013. Dr Alanazi is dedicated to imparting knowledge and particle skills to his students, fostering their understanding of transportation infrastructure, traffic control, and urban planning. Dr. Alanazi research interest include traffic flow modeling, intelligent transportation systems, sustainable transportation planning, and traffic safety. He published his research findings in reputable journals and he presented his work at national and international conferences.

Mohd Bilal received the B.Tech. degree in electrical engineering and the M.Tech. degree in instrumentation and control engineering from Aligarh Muslim University, Aligarh, India, in 2014 and 2016, respectively. He completed his Ph.D. degree in power system from Delhi Technological University, New Delhi, India in 2022. Currently, he is working as a postdoctoral research fellow in Department of Electrical Engineering Technology, University of Johannesburg, South Africa. He has published/presented several technical papers and conference proceedings in international journal of repute. His research interests include electric vehicles, renewable energy integration, artificial intelligence and optimization techniques.

AMMAR ARMGHAN (Senior Member, IEEE) received the bachelor's degree from COMSATS University, in 2006, the M.S. degree in electronics and communication engineering from the University of Nottingham, in 2010, and the Ph.D. degree from the Wuhan National Laboratory of Optoelectronic, Huazhong University of Science and Technology, Wuhan, China. He is currently an Associate Professor of electrical engineering with Jouf University. He has published more than 100 high-impact factor articles in the last three years. His research interests include machine learning, renewable energy, complementary metamaterial-based microwave and terahertz devices, biosensors, optics, and photonics. He is a Reviewer of several reputed journals.

MD REYAZ HUSSAN received his B.Tech. degree in 2014 and M.Tech degree in 2016 in Electrical Engineering and Instrumentation and Control, respectively from the Aligarh Muslim University, Aligarh, India. He is currently a Research Scholar in the Department of Electrical Engineering, Aligarh Muslim University. He is also currently a visiting Research Assistant with the department of Electrical Engineering, Qatar University, Doha, Qatar. His research interests include multilevel inverters and their control, photovoltaic systems, and multilevel inverter for solar PV applications.

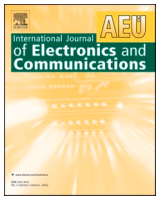
This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

Int. J. Electron. Commun. (AEÜ) 183 (2024) 155379

Contents lists available at ScienceDirect

International Journal of Electronics and Communications

journal homepage: www.elsevier.com/locate/aeue



A metasurface based close-proximity two-port circularly polarized MIMO antenna for mid-band sub-6 GHz 5G applications

Suresh Angadi [a,*], Yashna Sharma [a], N.S. Raghava [a], Thennarasan Sabapathy [b,c]

[a] Department of E.C.E, Delhi Technological University (DTU), Delhi 110042, India
[b] Advanced Communication Engineering (ACE), Centre of Excellence, Universiti Malaysia Perlis (UniMAP), Jalan Tiga, Pengkalan Jaya Business Centre, Kangar 01000, Perlis, Malaysia
[c] Faculty of Electronic Engineering & Technology, Universiti Malaysia Perlis (UniMAP), Pauh Putra Main Campus, 02600 Arau, Perlis, Malaysia

ARTICLE INFO

Keywords:
MIMO
Self-Isolation
Metasurface
New Radio frequency bands (n48, n79)
5G Spectrum

ABSTRACT

This work presents a compact dual-band two-port multiple input multiple output (MIMO) antenna with high performance that is specifically used for 5G applications. The proposed design comprises closely spaced, mirror-symmetrical X-shaped structures positioned at an edge spacing of 0.02 λ_0 between antenna elements. Subsequently, a metasurface is located at a distance of 0.27 λ_0 above the radiator to improve the performance of the MIMO antenna. The proposed antenna resonates at 3.27 GHz within a frequency range of (3.03–3.44) GHz in the primary band, achieving a peak realized gain of 7.57 dBi. In the secondary band, the resonance occurs at 4.78 GHz with a range of (4.01–5.87) GHz below −10 dB reflection coefficient and produces a peak realized gain of 5.60 dBi. After positioning the metasurface on top of the reference antenna, the peak realized gain improves by 116.28 % in the primary band and 69.69 % in the secondary band respectively. Whereas the isolation is increased to −30.01 dB from –22.26 dB in the primary band, similarly for the secondary band the isolation is amended to −21.00 dB from −18.20 dB. Moreover, circular polarization is realized at working frequencies after placing the metasurface on the reference antenna. Simulation and measurement results prove that the S-parameters exceed more than −20 dB in both frequency bands. Additionally, MIMO performance metrics, including envelope correlation coefficient (ECC), channel capacity loss (CCL), total active reflection coefficient (TARC), mean effective gain (MEG), and diversity gain (DG) are simulated and measured. The designed antenna proves suitable for the new radio (NR) 5G of n48 (3.55–3.70 GHz) in the primary band and the n79 (4.4–5.0 GHz) in the secondary band respectively.

1. Introduction

In modern wireless communication systems, multiple input multiple output (MIMO) technology is widely adopted to improve service quality, minimize latency, and satisfy for increasing demand for higher data rates [1,2]. The primary goal when designing a MIMO antenna is to minimize the correlation coefficient among received signals by various transmitting sections. It is crucial to achieve substantial isolation when the antenna elements are close to each other,this leads to prevent degradation in antenna performance, particularly concerning radiation patterns and interference [3]. To address these challenges, the literature discusses various methods for improving MIMO antenna isolation such as self-decoupling techniques [4,5], closely paired techniques [6,7], metamaterial based electromagnetic bandgap (EBG) structures [8], MTM absorber loading [9], metasurface reflector [10], and superstrate [11,12]. In [5,10,12,13] MIMO antennas face issues such as increased size and low gain with lesser antenna isolation. The performance metrics consider how the nearby positioning of antennas impacts the overall system performance. Also, when designing MIMO antennas, it's crucial to assess the envelope correlation coefficient (ECC) for both correlated and uncorrelated channel estimations for close proximity of antenna elements can lead to increased mutual coupling, which can degrade the performance by causing interference between the elements. The presence of nearby elements can affect the impedance matching of each antenna, potentially making it more difficult to achieve the desired impedance and leading to increased reflection and reduced power transfer. In this design, we have implemented a metasurface on top of the radiator to address these effects. Although the use of metasurface

* Corresponding author.
E-mail addresses: suresh.angadi.101@gmail.com (S. Angadi), thennarasan@unimap.edu.my (T. Sabapathy).

https://doi.org/10.1016/j.aeue.2024.155379
Received 28 April 2024; Accepted 4 June 2024
Available online 6 June 2024
1434-8411/© 2024 Elsevier GmbH. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

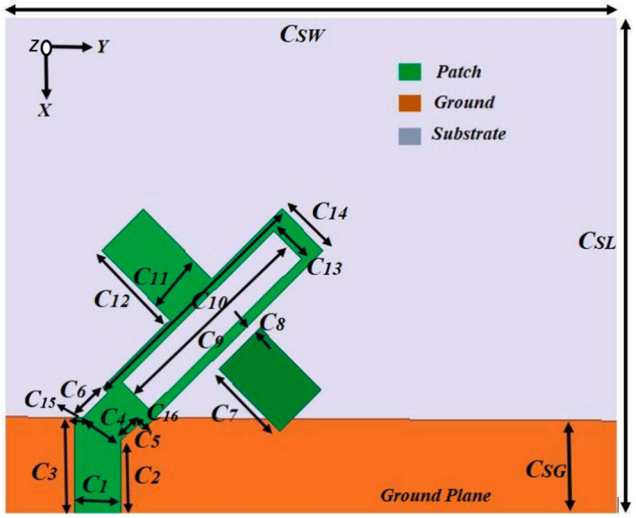


Fig. 1. The single-element antenna.

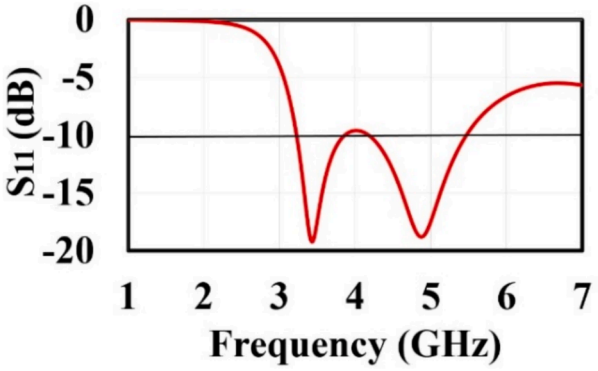


Fig. 2. Response of S-Parameter for a single element.

cannot entirely eliminate the constraints, it does provide a measure of protection in gain and isolation improvement in our design, mitigating these challenges to some extent. [14,15]. The reference [16] introduces a four-port MIMO antenna that exhibits good isolation. The primary purpose of the antenna is to provide coverage for the n48 band of Frequency Range-1 (FR-1) using Time Division Duplex (TDD) mode and used large dimensions. An additional decoupling structure is not required for port-to-port isolation in a simple MIMO antenna system, which leads to increased efficiency but a higher total antenna size [17].A connected ground MIMO antenna for 5G sub-6 GHz with four ports is built using a dielectric resonator. Dual bands were obtained by using the aperture feeding approach to excite the antenna [18].As a result, a simple X-shaped patch with two port arrangements using a metasurface proves that the designed antenna offers good isolation and enhanced gain even when the elements are close to each other Furthermore, the antenna design optimization ensures stable operation at the desired frequency. The proposed antenna projects superior isolation for closely spaced elements and the creation of a metasurface-based two-port antenna by determining the best design considerations using less space, and easy for fabrication.

The following are the primary contributions of this work:

In this work, a MIMO antenna is designed to operate at dual-band 5G frequencies and the two elements in this design are arranged in a mirror-symmetrical closely spaced configuration, featuring an edge-to-edge spacing of 1.7 mm. Moreover, the metasurface is created using an array of 2x3 unit cells, each comprising concentric circular rings. This metasurface is positioned atop with an air gap of 0.27 λ_0 produces that

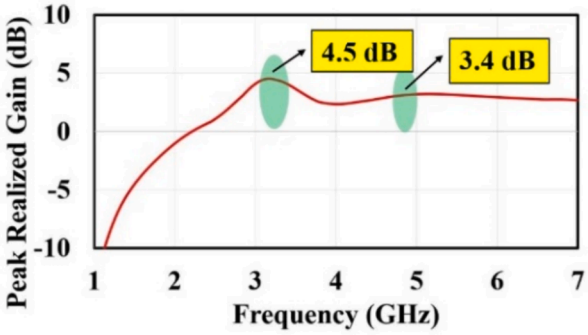


Fig. 3. Response of peak realized gain for single-element.

the polarization is converted, linear to circular polarization (CP) in both operating bands. After stacking the metasurface the peak realized gain is enhanced in the working bands. Furthermore, after optimizing the closely spaced design structure, the isolation improves to −30.01 dB from –22.26 dB in the primary band and to −21.00 dB from −18.20 dB in the secondary band. The configuration is further analyzed to examine its diversity performance in terms of MIMO configuration.

2. MIMO antenna geometry and design

2.1. Design of Single-Element antenna

Fig. 1 shows the arrangement of a single-port X-shaped structure resonating in dual-band and characterized by two distinct operating frequencies. Fig. 2 shows the response of the reflection coefficient resonating at 3.43 GHz ranging from (3.23–3.85 GHz) as the primary band and the secondary band resonates at 4.89 GHz ranging from (4.19–5.46 GHz). The peak realized gain for both resonances are 4.5 dBi and 3.4 dBi, which are shown in Fig. 3. Fig. 4. shows the responses of S-Parameters by varying the parameters C_{SG} and C_{10} and studying the effect of the proposed antenna for single-port configuration. The structure is designed and simulated using Ansys HFSS. The substrate used for simulation is Rogers 5880 with a dielectric constant of 2.2 and a loss tangent of 0.0009. The length and dimensions of the substrate are 23 x 23 mm^2 with a thickness of 1.6 mm. Table 1 shows the dimensions of a single-element antenna.

2.2. Evolution process and analysis

Fig. 5 presents the developmental stages of the proposed antenna and it is categorized into four distinct phases. The main objective is to design an antenna suitable for 5G applications. In stage 1 (ant-1), a basic rectangular patch is positioned diagonally from the feed, resonating at 5.99 GHz. Then, in ant-2 a slot is introduced within the inner part of the diagonally placed patch, resulting in a response at 5.34 GHz. To meet the specific application requirements we introduce slots to optimize the performance of a designed antenna.Ant-3 incorporates the addition of a stub from the opposite direction to the radiator, yielding a response at 5.14 GHz. In the final stage, an extra stub is introduced, transforming the structure into an X-shaped resonator operating in the dual-band for 5G applications.

2.3. Analysis of self-isolation with two-port MIMO

The design process of a two-port configuration includes the placement of a second element, which mirrors the same symmetrical structure of the first element and is positioned adjacent to it with a center-to-center distance (C_{DC}) of 13 mm between the antenna elements shown in Fig. 6. The gap between these two elements plays a pivotal role in assessing mutual coupling. In general, less mutual coupling occurs

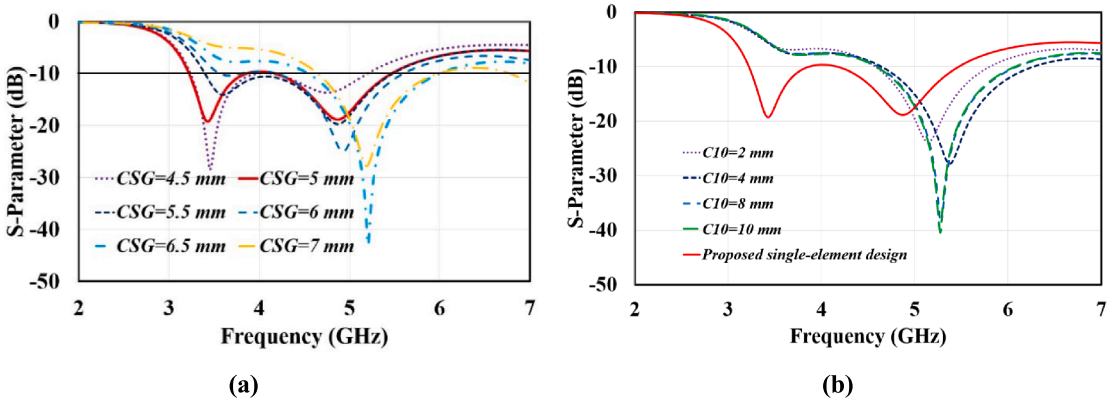


Fig.4. The effect of S_{11} by varying the parameters of single-element (a) C_{SG} (b) C_{10}.

Table 1
Parameter dimension of single-element antenna.

Parameter	C_{SL}	C_{SW}	C_{SG}	C_1	C_2	C_3
Value(mm)	23	23	5	2.4	3.8	5
Parameter	C_4	C_5	C_6	C_7	C_8	C_9
Value(mm)	2.2	1.2	1.7	4.5	1	11
Parameter	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15},C_{16}
Value(mm)	13	3	5	2	3	0.4

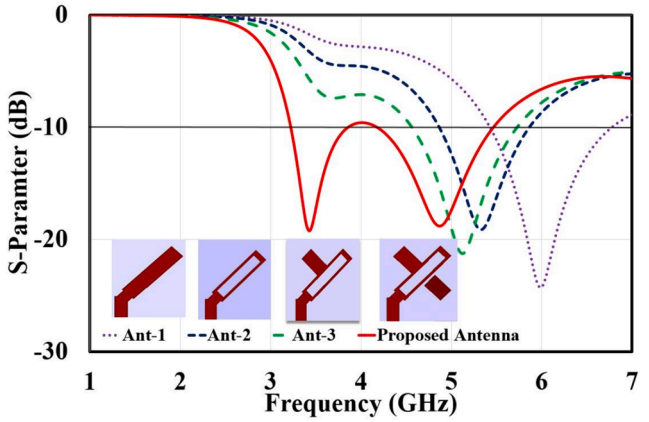


Fig. 5. Evolution steps of proposed single-element antenna.

between antenna elements when there is more space between them and it requires larger physical dimensions, but in this design, the elements are closely coupled with an edge spacing of 1.7 mm between antenna elements. This arrangement optimizes the design. Additionally, a ground strip (C_{DS}) is incorporated to regenerate the resonance observed in the basic antenna without using any decoupling circuit, and in the same way, the isolation is preserved without affecting the configuration. The analysis contrasts a two-port structure with and without the inclusion of a ground strip. As shown in Fig. 8, when the antenna is not using any ground strip, then the response of the reflection coefficient is shifted to the non-working band, and the peak realized gain of an antenna is reduced. The response presented in Fig. 7 reflects a two-port MIMO design with a common ground that effectively ensures minimum isolation of –22.26 dB in the primary band and −18.2 dB in the secondary band with a peak realized gain of 3.5 dBi and 3.3 dBi shown in Fig. 8(b). The separation between two antenna elements is $0.02\lambda_0$ at 3.2 GHz. Table 2 shows the parametric dimensions of the two-port antenna.

The S-Parameters and peak realized gain is studied by experimenting the edge-to-edge gap, (C_{DE}) with values of 5 mm, 3 mm,2.5 mm, 2

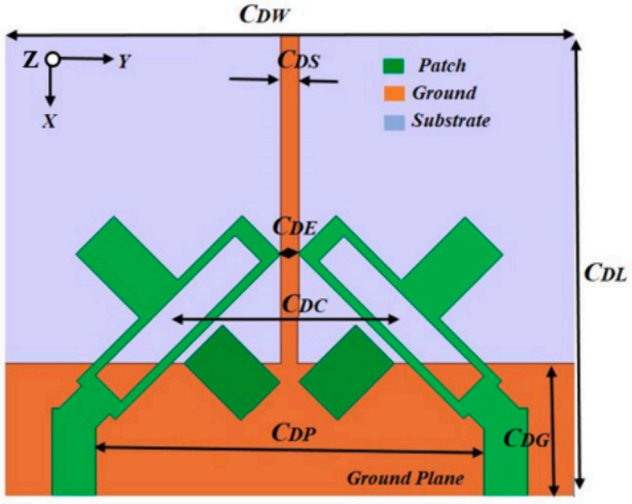


Fig. 6. Two-element MIMO Antenna.

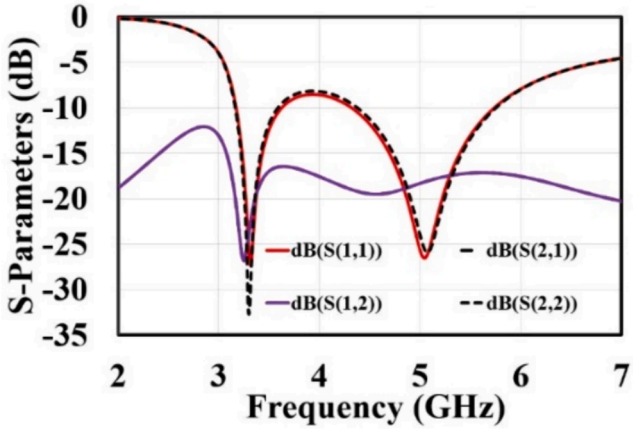


Fig. 7. Response of S-Parameter in two-port MIMO.

mm,1.7 mm and 1 mm and the results are presented in Figs. 9 and 10. Similarly, the parameter of C_{DG} is adjusted and optimized based on the response shown in Fig. 9(a).When C_{DE} increases, the substrate size concurrently enlarges. It can be seen that in the primary band, the reflection coefficient and transmission coefficient experience an increase while in the secondary band, it undergoes a decrease, as illustrated in Fig. 10. Based on this parametric analysis, we optimized the

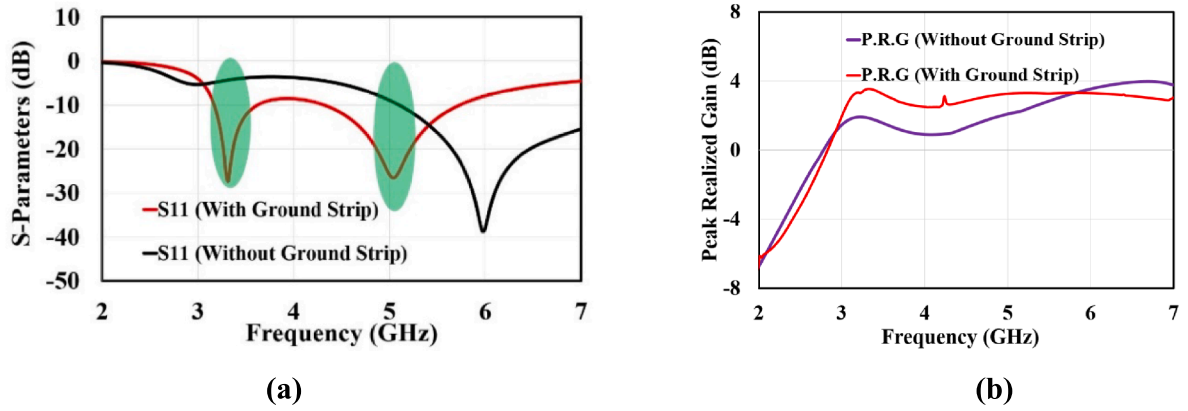


Fig. 8. Comparison of two-port structure with and without ground strip (a) Response of S-parameters. (b) Peak realized gain vs frequency.

Table 2
Parametric dimensions of two-port antenna.

Parameter	C_{DW}	C_{DL}	C_{DS}	C_{DE}	C_{DC}	C_{DP}	C_{DG}
Units(mm)	*30*	*23*	*1.7*	*1.7*	*13*	*22.4*	*7.5*

performance of the antenna structure for our specific application. Finally, a center-to-center distance (C_{DE}) of 1.7 mm between the antenna elements was selected for superior results compared to other parameters.

3. Developing and analyzing a decoupling circuit for optimal performance

3.1. Design of metasurface

In order to enhance the performance of the MIMO antenna, a metasurface-superstrate layer is incorporated into the design. This addition aims to mitigate mutual coupling in both the lower and upper bands of the closely coupled antenna. The metasurface array eventually helped to improve the antenna gain while achieving polarization conversion at operating frequencies. The design procedure of the single-

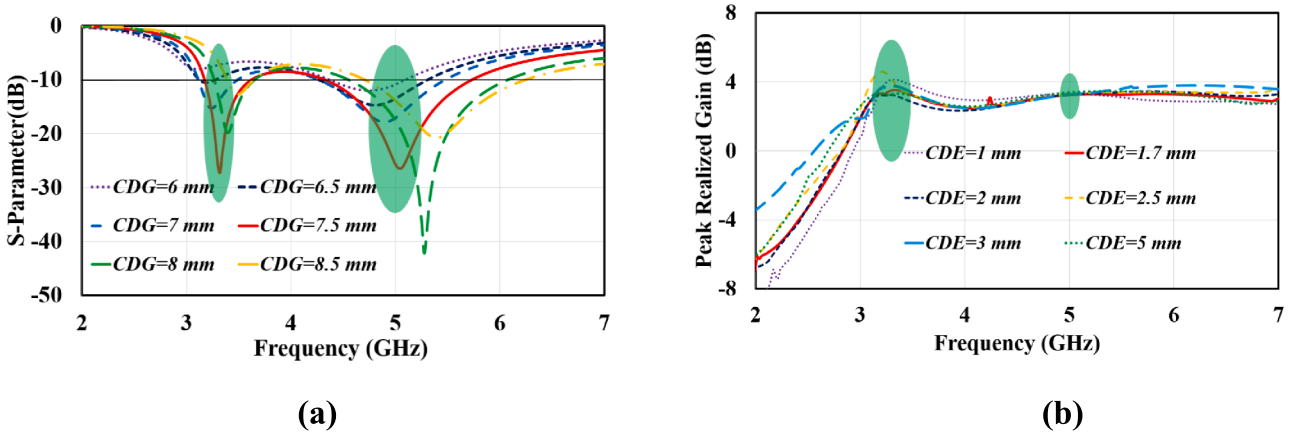


Fig. 9. The Effect of S_{11} and P.R.G by varying the parameters for Two-port antenna **(a)** C_{DG} **(b)** C_{DE}.

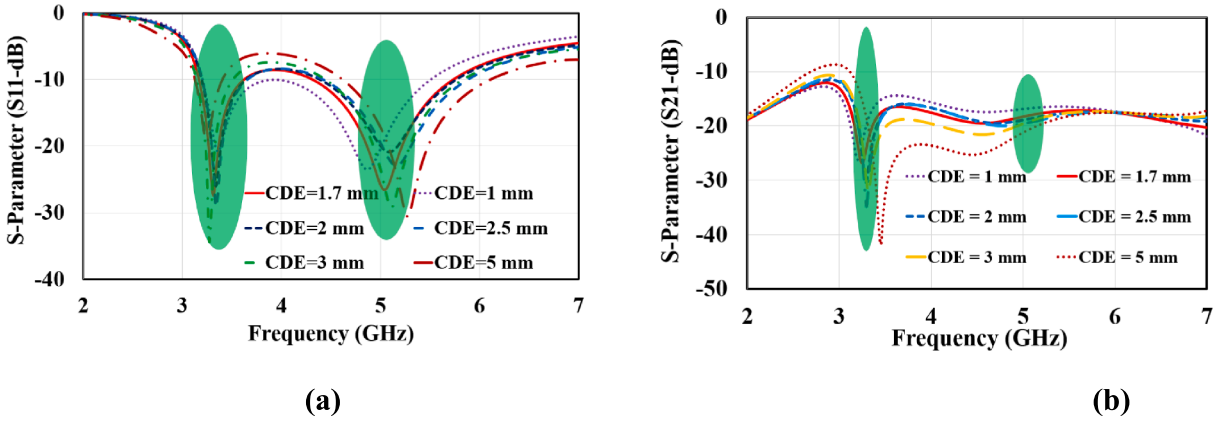


Fig. 10. The effect of parameter C_{DE} (a) S_{11} vs frequency. (b) S_{21} vs frequency.

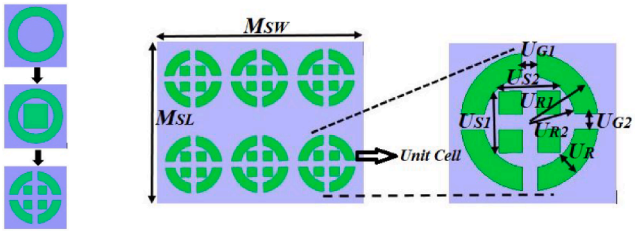


Fig. 11. Unit cell: **(a)** Formation of unit cell **(b)** Unit cell structure and its geometry.

element unit cell is shown in Fig. 11 (a). Similarly, an array of 2x3 unit cells is positioned on the identical substrate (layer-2) utilized in the reference antenna shown in Fig. 11 (b). The unit cell simulation analysis is carried out by employing periodic boundaries and floquet ports are shown in Fig 12 (a). Fig 12(b) illustrates the phase variance between the incident and reflected waves of the proposed antenna. The graph demonstrates that the antenna's reflection phase remains consistent within the operational bandwidth, ranging between + or – 90 degrees. Table 3 shows the dimensions of the proposed metasurface structure.

3.2. MIMO system integrated with metasurface

The primary motivation for our work is to enhance isolation between closely spaced antenna elements in a MIMO system. We have achieved this by positioning the antenna elements in mirror-symmetrical structures opposite each other. Based on our extensive literature review, we selected a metasurface as a decoupling circuit. This choice was driven by its effectiveness in reducing mutual coupling and enhancing the peak realized gain of closely coupled elements. References [3] and [15] review the various mutual coupling reduction techniques for MIMO antennas, which provided a foundational understanding of our design. We compared our proposed design with references [4 to 13], which focus on different approaches to improve isolation and performance in MIMO systems. The comparative analysis with existing works highlights the effectiveness of our approach.

3.2.1. Design stages and their Descriptions

The following Table 4 summarizes the key stages of our antenna design and the improvements achieved at each stage.

The proposed metasurface array reduces the coupling effect of the closely coupled antenna elements by generating reflected waves. The superstrate is placed on top of the MIMO antenna with an air gap of 25 mm ($0.26\lambda_0$) represented in the two-layer of Fig. 13. The metasurface helps in improving the peak realized gain from 3.50 dB to 7.57 dB in the

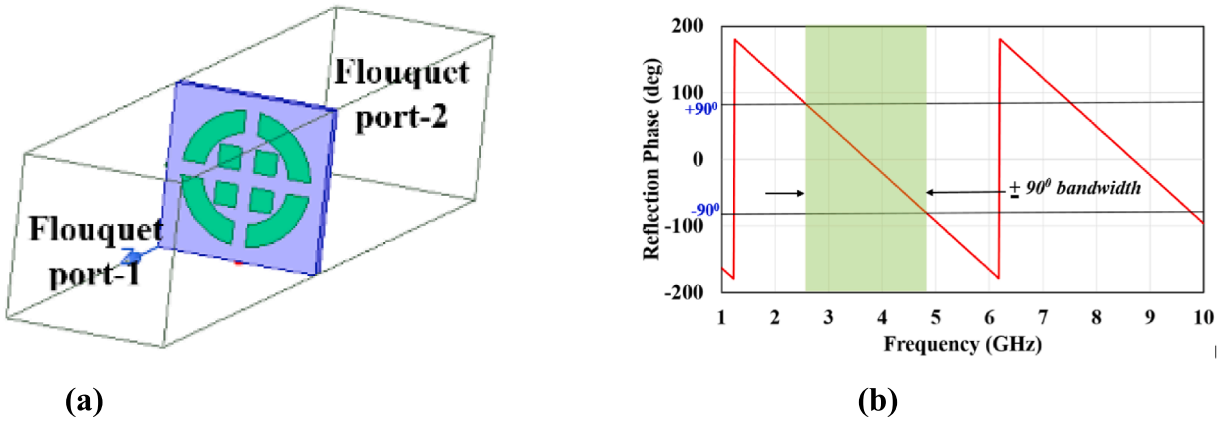


Fig. 12. Simulation and the response of the unit cell.**(a)** Unit cell excitation. **(b)** Reflection phase in degrees.

Table 3
Parametric dimension of unit cell and metasurface.

Parameter	H_1	H_2	H_3	M_{SW}	M_{SL}	U_{G1}	U_{G2}	U_{S1}	U_{S2}	U_{R1}	U_{R2}	U_R
Value(mm)	1.6	25	1.6	30	23	1	1	4	4	4.5	3	1.5

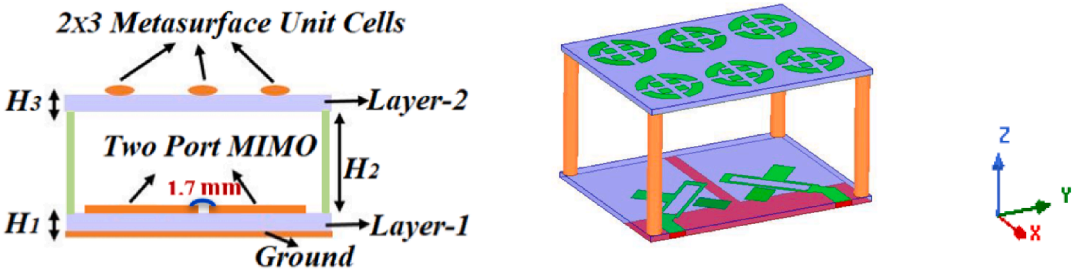


Fig. 13. **(a)** Equivalent model of proposed structure in side-view. **(b)** 3D-View of the proposed design.

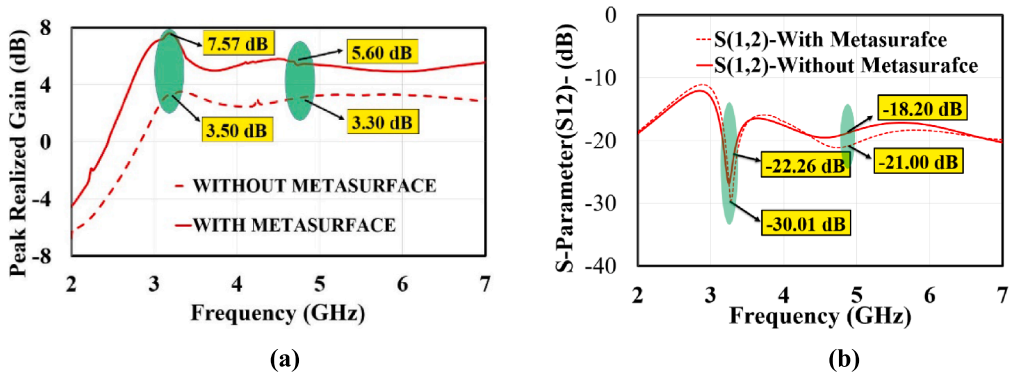


Fig. 14. Simulation response of with and without metasurface: **(a)** Peak realized gain vs frequency. **(b)** Isolation (S_{21}) vs frequency.

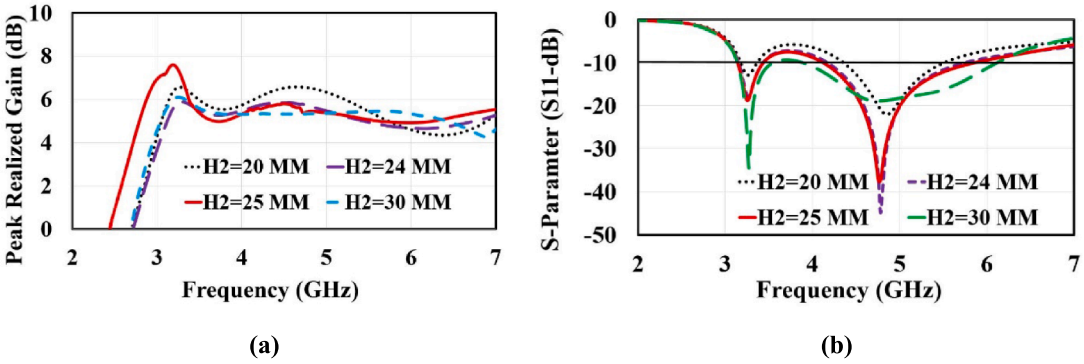


Fig. 15. Simulated characteristics of the proposed antenna with a variation of different metasurface heights(H2): **(a)** Peak realized gain vs frequency. **(b)** S-parameters vs frequency.

Table 4
Performance comparision of proposed structure in stage-wise.

Stage	Description	Improvements
Stage 1	Initial Antenna Design	The basic single-port antenna design is extended to two-port antenna elements is positioned in mirror symmetry. No decoupling technique applied.
Stage 2	Implementation of Metasurface	A metasurface has been introduced as a decoupling circuit and is positioned 25 mm above the radiator.
Stage 3	Optimization of Metasurface Parameters	Fine-tuning the metasurface dimensions and substrate height to achieve optimal isolation and gain.
Stage 4	Comparative Analysis with Previous designs (References. [4] to [13])	Evaluation of the proposed design against previous works, demonstrating superior isolation and peak realized gain.

primary band and in the secondary band the gain increases from 3.30 dB to 5.60 dB are shown in Fig. 14(a). The isolation experiences a notable enhancement in the primary band, escalating from –22.26 to −30.01 dB at 3.2 GHz. Conversely, in the secondary band, the isolation rises from −18.20 dB to −21.00 dB at 4.78 GHz as shown in Fig. 14(b). After stacking the metasurface with different heights (H_2), the peak realized gain and isolation are observed through simulation as shown in Fig. 15. Finally layer 2 is fixed with an air gap of 25 mm from the radiator, and also a circular polarization is realized in the operating band assuming metasurface acts as a polarization converter [19,20] Table 4.

3.3. Polarization conversion metasurface

The main purpose of the metasurface's unit cell elements is to alter the properties of electromagnetic waves. In this work, the metasurface improves the gain and isolation of the planned antenna and permits polarization conversion. The reference antenna proposes linear polarization when it is not equipped with a metasurface. After integrating the metasurface on the reference antenna with a free space height of 25 mm, circular polarization is acquired at 3.2 GHz and 4.2 GHz, as seen in Fig. 16(a), By varying the height (H2) of the metasurface, an analysis is carried out to assure that the suggested antenna functions in a circularly polarized mode. The circular polarization projection is used to determine the height of the metasurface (layer-2) through the iterations depicted in Fig. 16(b). Finally, the maximal realized gain and isolation are enhanced by stacking the metasurface above the radiator, as mentioned in section 3.2.

3.4. Surface current distribution

The movement of electrical current across the surface of an antenna structure is referred to as surface current distribution. Determining the antenna's polarization is largely dependent on this dispersal. An antenna will produce both electric and magnetic fields when subjected to an electromagnetic wave. Meanwhile, for phases (0^0,90^0 and 180^0) of the suggested antenna, the surface current distribution is shown at 3.2 GHz. On the antenna's surface, these fields induce current flow. The surface current flow for antennas 1 and 2 is in a clockwise direction (LHCP), as seen in Fig. 17. The geometry and electrical characteristics of the antenna control how these currents are distributed.

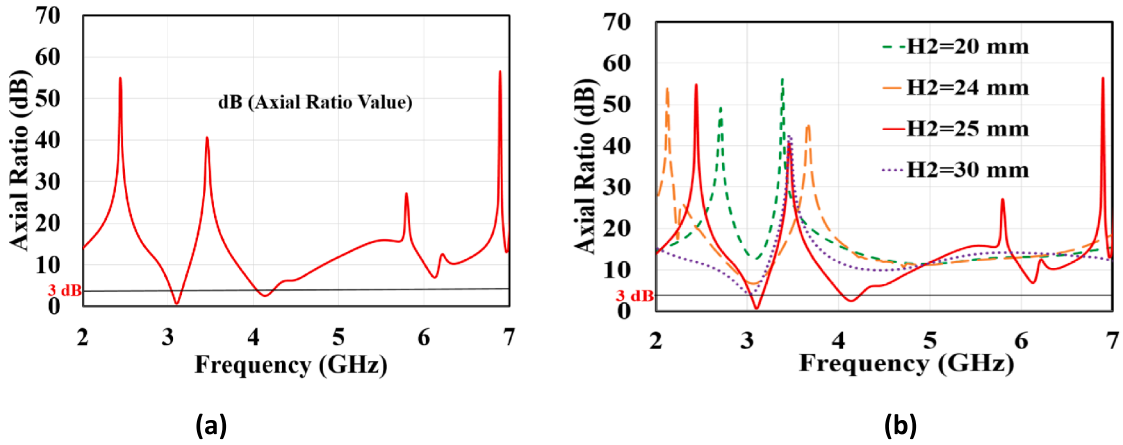


Fig. 16. Axial ratio analysis: (a) Axial ratio vs frequency. (b) Response of axial ratio with different values of H_2.

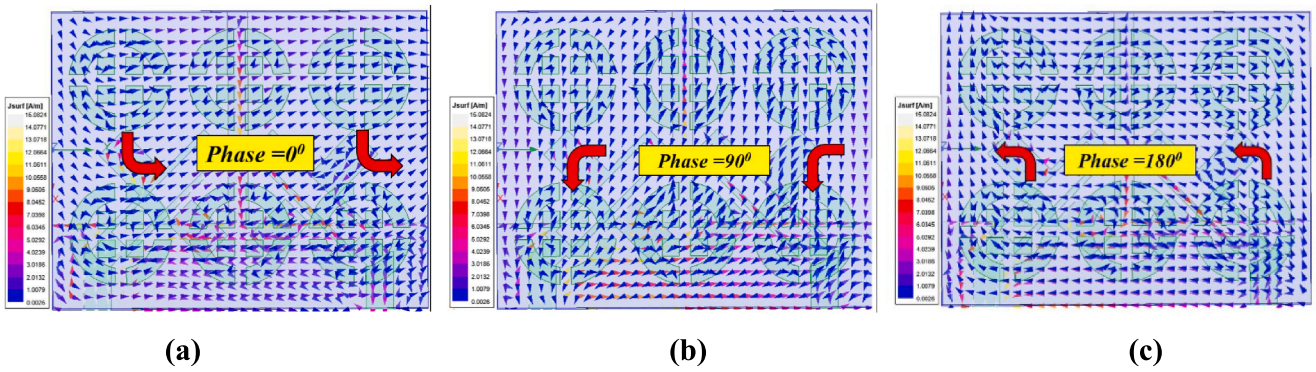


Fig. 17. Surface current distribution at 3.2 GHz.

Table 5
Comparison of single element, reference antenna and proposed structures.

Antenna parameters	Single element		Reference antenna		Proposed antenna	
Size (L x W)	23 X 23 mm^2		23 X 30 mm^2		23 X 30 mm^2	
Resonant frequencies (F_r)	3.43 GHz	4.89 GHz	3.31 GHz	5.05 GHz	3.25 GHz	4.78 GHz
Return Loss(dB)	−19.24 dB	−18.81 dB	−27.37 dB	−26.54 dB	−18.73 dB	−37.45 dB
Fractional bandwidth	17.51 %	26.32 %	13.52 %	28.45 %	9.11 %	35.035 %
Isolation	Not applicable		−22.26 dB	−18.20 dB	−30.01 dB	−21.00 dB
Peak realized gain(dB)	4.5 dBi	3.4 dBi	3.50 dBi	3.30 dBi	7.57 dBi	5.60 dBi

4. Results and discussion

4.1. Performance comparison table for single-element, two element and proposed structure with metasurface

For the 5G New Radio spectrum, the 3rd Generation Partnership Project (3GPP) has developed Frequency Division Duplex (FDD) frequency bands, notably in the FR1 range [1]. In this setup, our proposed design effectively fulfills the operational requirements of NR bands n48 and n79 respectively. In Table 5, we present a performance comparison between single-port and two-port design structures. Initially, the configuration was developed for a single element, later it extended to a

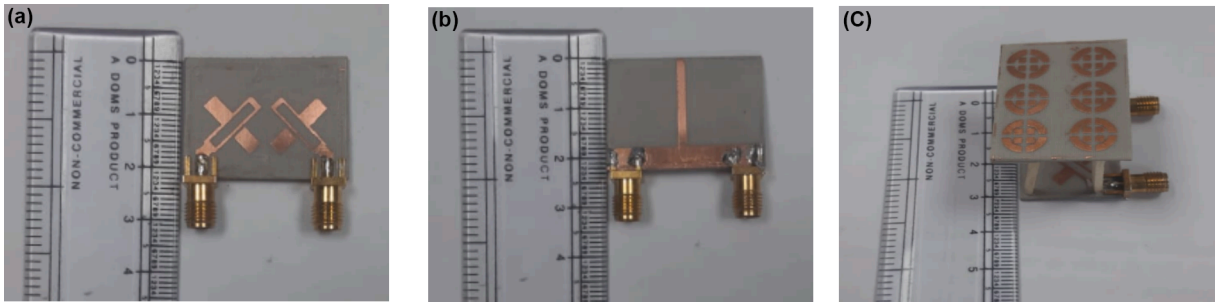


Fig. 18. Proposed MIMO Antenna after Fabrication Process (**a**) Top view (**b**) Back view (**c**) Top-view with metasurface.

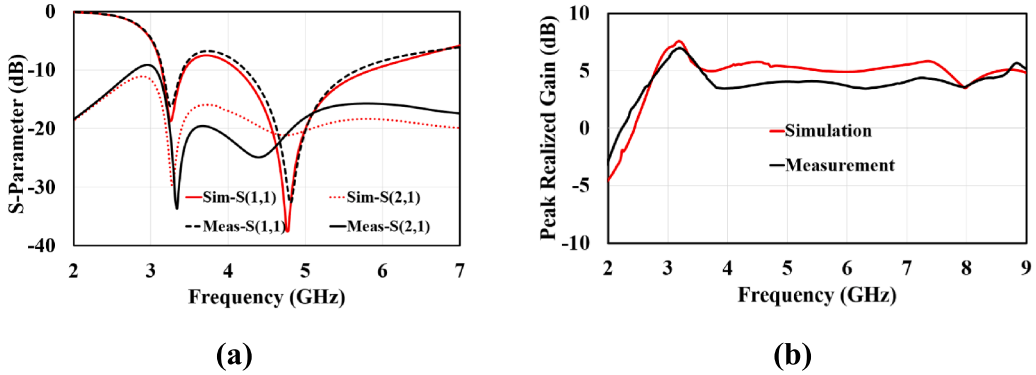


Fig. 19. Comparision of simulation and measured results: (a) S-parameters Vs frequency.

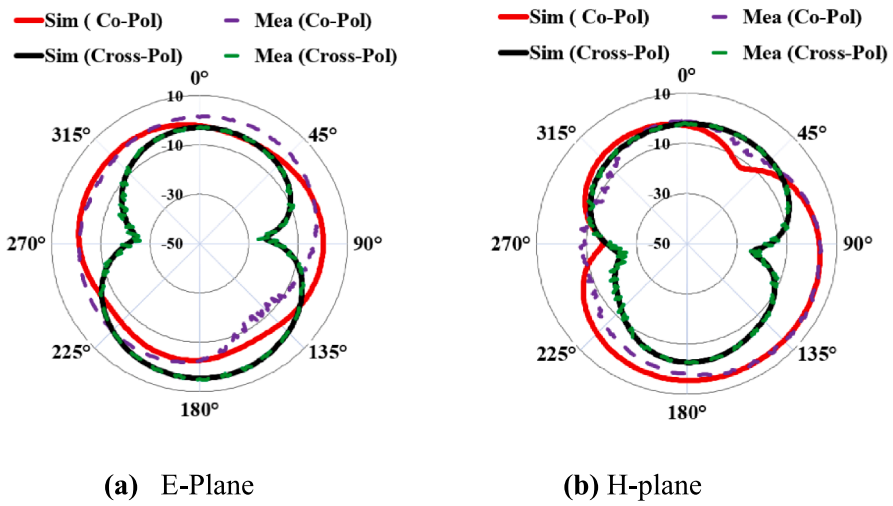


Fig. 20. Co & Cross-Polarisations at 3.27 GHz.

1x2 MIMO configuration operating in dual-band mode. The key observation from the proposed design is it includes a narrow impedance bandwidth in the primary band compared to a wider bandwidth in the secondary band. Particularly, the maximum peak realized gain increases by integrating the metasurface-superstrate into the design. The isolation between closely coupled elements appears as a critical factor. Through optimization of design parameters and the inclusion of metasurface, isolation is enhanced, ensuring the proposed design can be operated satisfactorily for the 5G spectrum.

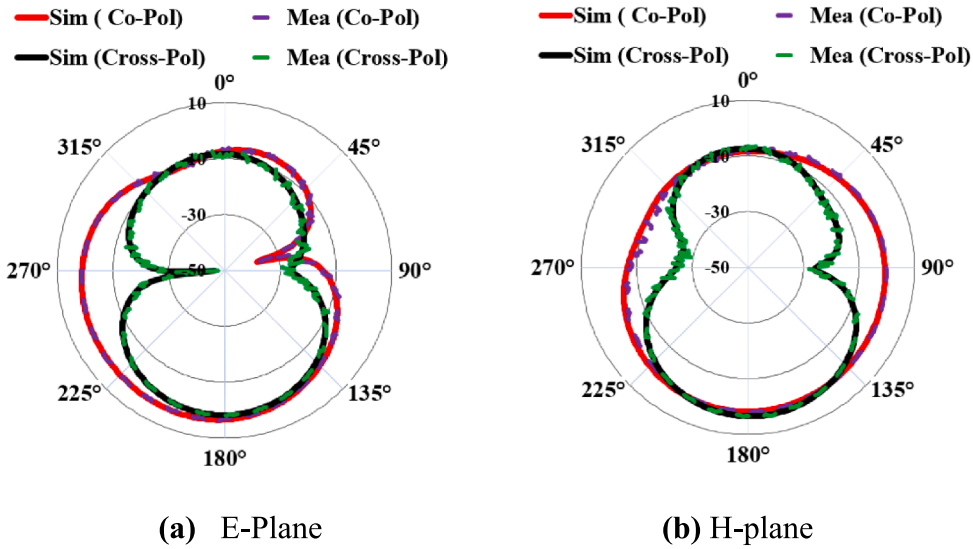


Fig. 21. Co & Cross-Polarisations at 4.78 GHz.

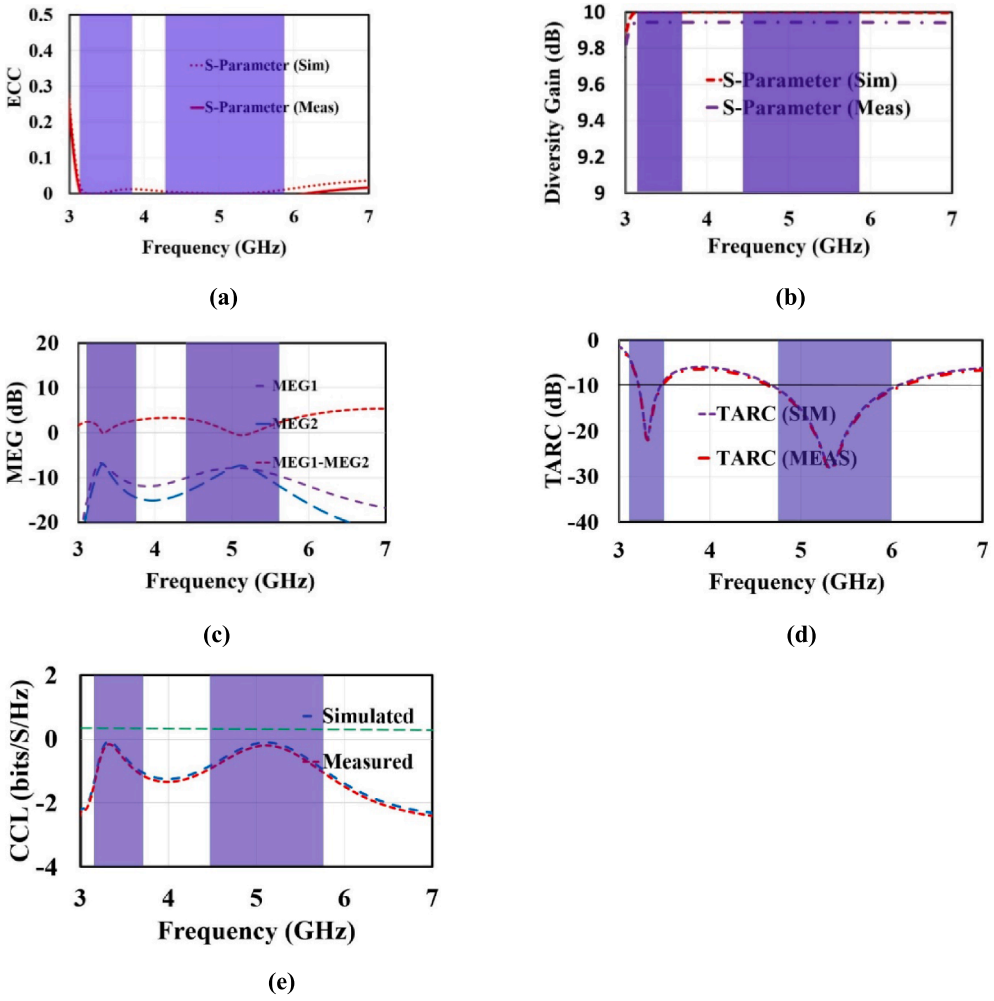


Fig. 22. Response of simulation and measurement results of MIMO Parameters: (a) ECC (b) Diversity Gain (C) TARC (d) MEG (e) CCL.

4.2. S-parameters and peak realized gain

Fig. 18. Shows the fabricated images of the proposed configuration. In Fig. 19, the responses of simulated and measured results of original antenna parameters are illustrated. The measured S_{11} exhibits a bandwidth of 600 MHz and a maximum isolation of −30.01 dB in the primary band. Where in the secondary band, the measured bandwidth is approximately 1.75 GHz, with an isolation of −21.00 dB are shown in Fig. 19(a). Moreover, the measurement of the peak realized gain of port-1 is determined by connecting port-2 to a 50 Ω matched load, revealing an approximate value of 7.57 dBi in the primary band and 5.60 dBi in the secondary band, due to the fabrication losses, an estimation of 1 dB of peak realized gain is reduced in the secondary band reflected in Fig. 19 (b).

4.3. Radiation pattern

The distribution of radiated power as a function of direction in space is represented graphically by an antenna's radiation pattern. It offers vital details regarding the energy radiation of an antenna and is necessary to comprehend the functionality and appropriateness of the antenna.

The anechoic chamber gathers the antenna's far-field data. The proposed antenna is used as a receiver (R_X), while a conventional gain horn antenna serves as the transmitter (T_X). To obtain measurements from various orientations, the antenna under test is rotated. Figs. 20 and 21 show the simulated and measured results of the radiation patterns for the proposed closely coupled metasurface-based two-port MIMO antenna for both operating frequencies, it is observed that the co-polarization of the E-plane pattern is omnidirectional and the cross-polarization is bi-directional whereas in the H-plane the cross-polarisation is observed as omni-directional and co-polarisation is in bi-directional resembling like figure of eight representation.

4.4. MIMO parameters

The measured and simulation values of the reflection coefficient and transmission coefficient isolation metrics for the designed antenna consistently adhere to and remain below the standard values for MIMO parameters [21]. The envelope correlation coefficient (ECC) of the proposed design is displayed in Fig. 22(a). The measured and simulated ECC values in the working bands are less than 0.01. Similarly, Fig. 22(b) shows the diversity gain (DG), which is calculated using the expression, $DG = 10\times\sqrt{(1 - ECC^2)}$. The calculated diversity gain (DG) values show acceptable MIMO antenna performance since they are within the typical value range. Fig. 22(c) displays the mean effective gain (MEG) responses. The following expression is used to calculate MEG. $MEG_i = 0.5(1\text{-}\sum_{j=1}^{k}(P_{ij}))$ where k is no. of antennas which satisfies the condition less than 3 dB [22]. The total active reflection coefficient (TARC) response are shown in Fig. 22(d) and it is calculated using the expression $\gamma_a^t = \sqrt{\frac{(S_{11}+S_{12}e^{j\theta})^2+(S_{21}+S_{22}e^{j\theta})^2}{2}}$, and the response is maintained below 0 dB. The calculation of channel capacity loss (CCL) is determined by utilizing the

Table 6
Comparison of proposed design with previous works.

Ref. No	Size (mm x mm)	Decoupling Circuit	Center Frequency (Fc)	No. of operating bands	Maximum Isolation	Edge Spacing	LP/ CP	Antenna Bandwidth	ECC	Peak Realized Gain
4	150x75	Self-Decoupled	3.75 GHz	1	−10.5 dB	70 mm	LP	3.3–4.2 GHz	0.08	NG
5	150X75	Self-Decoupled	3.5 GHz, 4.9 GHz	2	−17.5 dB, −20 dB	31.2 mm	LP	(3.4–3.6), (4.8–5.0)	0.14, 0.12	NG
6	90 x 100	MTM-EMBG	5.6 GHz	1	−17 dB	20 mm	LP	8–9.25	NG	7 dBi
7	67.41x32.82	MTM-DS	10 GHz	1	−38 dB	20.04 mm	LP	9.56–10.63	NG	5.4 dBi
8	119 x 119	Double-layer Mushroom Structure	GHz		−42 dB	63 mm		−2.45		5.1 dBi
9	24.4x32.2	Metamaterial Absorber	5.5 GHz	1	−34.77 dB	9 mm	LP	–	0.05	7.74 dBi
10	83.8x83.8	Metasurface Reflector	2.4 GHz	1	−14.5 dB	18.4 mm	CP	2.23–2.91	NG	7.02 dBi
11	56x38	Metamaterial Superstrate	5.7 GHz	1	−41 dB	2.2 mm	NG	5.5–5.92	0.26	7.4 dBi
12	60.6x48.5	Metasurface Reflector	3.7 GHz, 4.1 GHz	2	−28.4 dB, −25.9 dB	2.8 mm	LP	NG	0.01	8.5 dBi, 7.9 dBi
13	52x77.5	Slots on Ground Plane	2.4 GHz, 5.2 GHz	2	−15 dB	12 mm	LP	2.4–2.48, 5.18–5.84	0.2	7.5 dBi
This Work	**23x30**	**Metasurface- Reflector**	**3.27 GHz, 4.78 GHz**	**2**	**−30.01 dB, −21.00 dB**	**1.7 mm**	**CP**	**3.14-3.74, 4.0—5.87**	**0.01**	**7.57 dBi, 5.6 dBi**

LP/CP = Linear/Circular polarization, ECC = Envelope correlation coefficient, NG = Not given, MTM EMBG = Metamaterial Elecro- magnetic bandgap, MTM-DS = Metamaterial decoupling slab.

provided expression. $CCL = -log_2 Det(K^R)$ where $K^R = \begin{bmatrix} \varnothing_{11} & \varnothing_{12} \\ \varnothing_{21} & \varnothing_{22} \end{bmatrix}$ the elements of the correlation matrix are calculated by $\varnothing_{ii} = (1 - \left|\sum_{n=1}^{2}(S^*_{ij}.S_{nj})\right|)$ and $\varnothing_{ij} = -\left|\sum_{n=1}^{2}(S^*_{ij}.S_{nj})\right|$ where i and j are 1 or 2 according to the port number [20]. It is evident that Fig. 22(e) indicates that the channel capacity loss (CCL) values consistently remain within the acceptable limit, as the allowed threshold for CCL is set at < 0.4 bits/s/Hz Table 6.

5. Conclusion

This work introduces a closely coupled two-port MIMO using polarization conversion metasurface is designed for NR band of n48 and n79 operating in 5G applications. Initially, a single antenna element is transformed into a two-port MIMO. To enhance the performance of the proposed MIMO system, a metasurface with identical dielectric properties to the radiator is integrated and placed on top of the radiator. With these the isolation and peak realized gain was improved. The proposed metasurface based two-port MIMO antenna produces a circular polarization on operating bands. Operating in circular polarization mode the surface current distribution is projected and analyzed for two elements in the design. The diversity performance metrics of the proposed antenna are ECC, MEG, CCL,TARC and DG as evaluated and compared with measured results for proposed configuration. Thus, it can be concluded that the proposed MIMO antenna is well-suited for next-generation 5G technology.

CRediT authorship contribution statement

Suresh Angadi: Writing – original draft, Validation, Methodology, Conceptualization. **Yashna Sharma:** Validation, Supervision, Resources. **N.S. Raghava:** Supervision, Resources, Formal analysis. **Thennarasan Sabapathy:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [Suresh Angadi reports a relationship with Delhi Technological University that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personalrelationships that could have appeared to influence the work reported in this paper.].

Data availability

No data was used for the research described in the article.

References:

[1] Lin X. An Overview of 5G Advanced Evolution in 3GPP Release 18. IEEE Comm Stand Mag Sep. 2022;6(3):77–83. https://doi.org/10.1109/MCOMSTD.0001.2200001.

[2] Gupta A, Jha RK. A Survey of 5G Network: Architecture and Emerging Technologies. IEEE Access 2015;3:1206–32. https://doi.org/10.1109/ACCESS.2015.2461602.

[3] Chen X, Zhang S, Li Q. A Review of Mutual Coupling in MIMO Systems. IEEE Access 2018;6:24706–19. https://doi.org/10.1109/ACCESS.2018.2830653.

[4] Sun L, Li Y, Zhang Z, Wang H. Self-Decoupled MIMO Antenna Pair With Shared Radiator for 5G Smartphones. IEEE Trans Antennas Propagat May 2020;68(5):3423–32. https://doi.org/10.1109/TAP.2019.2963664.

[5] Ren Z, Zhao A. Dual-Band MIMO Antenna With Compact Self-Decoupled Antenna Pairs for 5G Mobile Applications. IEEE Access 2019;7:82288–96. https://doi.org/10.1109/ACCESS.2019.2923666.

[6] Alibakhshikenari M, et al. Study on isolation improvement between closely-packed patch antenna arrays based on fractal metamaterial electromagnetic bandgap structures. IET Microwaves Antennas Propag Nov. 2018;12(14):2241–7. https://doi.org/10.1049/iet-map.2018.5103.

[7] Alibakhshikenari M, et al. Interaction Between Closely Packed Array Antenna Elements Using Meta-Surface for Applications Such as MIMO Systems and Synthetic Aperture Radars. Radio Sci Nov. 2018;53(11):1368–81. https://doi.org/10.1029/2018RS006533.
[8] Zhai G, Chen ZN, Qing X. Enhanced Isolation of a Closely Spaced Four-Element MIMO Antenna System Using Metamaterial Mushroom. IEEE Trans Antennas Propagat Aug. 2015;63(8):3362–70. https://doi.org/10.1109/TAP.2015.2434403.
[9] Garg P, Jain P. Isolation Improvement of MIMO Antenna Using a Novel Flower Shaped Metamaterial Absorber at 5.5 GHz WiMAX Band. IEEE Trans Circuits Syst II Apr. 2020;67(4):675–9. https://doi.org/10.1109/TCSII.2019.2925148.
[10] Ameen M, Ahmad O, Chaudhary RK. Wideband circularly-polarised high-gain diversity antenna loaded with metasurface reflector for small satellite applications. Electron Lett Jul. 2019;55(15):829–31. https://doi.org/10.1049/el.2019.1645.
[11] Mark R, Singh HV, Mandal K, Das S. Mutual coupling reduction using near-zero ε and μ metamaterial-based superstrate for an MIMO application. IET Microwaves Antennas Propag May 2020;14(6):479–84. https://doi.org/10.1049/iet-map.2019.0382.
[12] Niu Z, Zhang H, Chen Q, Zhong T. Isolation Enhancement in Closely Coupled Dual-Band MIMO Patch Antennas. Antennas Wirel Propag Lett Aug. 2019;18(8):1686–90. https://doi.org/10.1109/LAWP.2019.2928230.
[13] Deng J, Li J, Zhao L, Guo L. A Dual-Band Inverted-F MIMO Antenna With Enhanced Isolation for WLAN Applications. Antennas Wirel Propag Lett 2017;16:2270–3. https://doi.org/10.1109/LAWP.2017.2713986.
[14] Sharawi MS. Printed Multi-Band MIMO Antenna Systems and Their Performance Metrics [Wireless Corner]. IEEE Antennas Propag Mag Oct. 2013;55(5):218–32. https://doi.org/10.1109/MAP.2013.6735522.
[15] Nadeem I, Choi D-Y. Study on Mutual Coupling Reduction Technique for MIMO Antennas. IEEE Access 2019;7:563–86. https://doi.org/10.1109/ACCESS.2018.2885558.
[16] Upadhyaya T, et al. Quad-port MIMO antenna with high isolation characteristics for sub 6-GHz 5G NR communication. Sci Rep Nov. 2023;13(1):19088. https://doi.org/10.1038/s41598-023-46413-4.
[17] Pant A, Singh M, Parihar MS. A frequency reconfigurable/switchable MIMO antenna for LTE and early 5G applications. AEU-Int J Electron C Mar. 2021;131:153638. https://doi.org/10.1016/j.aeue.2021.153638.
[18] Upadhyaya T, et al. Aperture-Fed Quad-Port Dual-Band Dielectric Resonator-MIMO Antenna for Sub-6 GHz 5G and WLAN Application. Int J Antennas Propag Aug. 2022;2022:1–13. https://doi.org/10.1155/2022/4136347.
[19] Loncar J, Grbic A, Hrabar S. A Reflective Polarization Converting Metasurface at ${X}$ -Band Frequencies. IEEE Trans Antennas Propagat Jun. 2018;66(6):3213–8. https://doi.org/10.1109/TAP.2018.2816784.
[20] Angadi S, Sharma Y, Raghava NS, Sabapathy T. A low profile circularly polarized metasurface-based ultra-wideband 4x4 MIMO antenna for 5G NR band FR2 frequencies. AEU-Int J Electron C May 2024;178:155282. https://doi.org/10.1016/j.aeue.2024.155282.
[21] Aghoutane B, Das S, El Ghzaoui M, Madhav BTP, El Faylali H. A novel dual band high gain 4-port millimeter wave MIMO antenna array for 28/37 GHz 5G applications. AEU-Int J Electron C Feb. 2022;145:154071. https://doi.org/10.1016/j.aeue.2021.154071.
[22] Nielsen JO, Pedersen GF, Olesen K, Kovacs IZ. Computation of mean effective gain from 3D measurements. 1999 IEEE 49th Vehicular Technology Conference (Cat. No.99CH36363) Houston, TX, USA: IEEE 1999:787–91. https://doi.org/10.1109/VETEC.1999.778304.

2024 IEEE 9th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 5-7, 2024

2024 IEEE 9th International Conference for Convergence in Technology (I2CT) | 979-8-3503-9447-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/I2CT61223.2024.10543334

A Novel Enhancement Mode GaN HEMT with MIS and p-GaN structure

1st Tanvika Garg
Electronics and Communication Engineering Department
Delhi Technological University
New Delhi, India
tanvikagarg_2k21phdec06@dtu.ac.in

2nd Sumit Kale
Electronics and Communication Engineering Department
Delhi Technological University
New Delhi, India
sumit.kale@dtu.ac.in

***Abstract*—This paper reports a novel enhancement mode GaN HEMT designed to mitigate gate leakage current. The device combines a p-GaN layer with HfO_2 layer beneath the metal. The p-GaN layer acts as a barrier, diminishing tunneling, while HfO_2 exhibits fewer trap sites compared to SiN, resulting into reduced gate leakage current. Moreover, the proposed device achieves improved gate leakage current, maximum drain current and threshold voltage compared to the devices in the already reported literature.**

***Index Terms*—p-GaN, HfO_2 and Gate leakage**

Introduction

GaN HEMTs have revolutionized semiconductor devices, finding applications in power electronics [1] [2], wireless communications, and radar systems. However, gate leakage current in GaN HEMTs is a major concern for power electronics applications, causing increased power dissipation and reduced switching efficiency, limiting the performance and reliability of these devices. Overcoming these challenges is essential for advancing the effectiveness of GaN HEMTs in power electronics.

In the past decade, various studies were reported to improve the gate leakage current of the device. In some published reports [3]–[5], the gate leakage current was improved by: (i) including Al_2O_3/AlN composite gate insulator in p-GaN MISHEMT, (ii) employing Si-rich and stoichiometric SiN_x first passivation in buffer free GaN MISHEMT, and (iii) using vertically-ordered h-BN in AlGaN/GaN MISHEMT on Si substrate. Nevertheless, the above-mentioned techniques are not optimal for GaN MISHEMT. The first method exhibited higher capacitance than the conventional p-GaN MISHEMT leading to lower switching speed of the device. Further, the barrier leakage current is increased by using the second method. Moreover, in the third technique, a series capacitance is introduced due to which the threshold voltage is shifted to more negative. Hence, further improvements are required to ensure that performance remains uncompromised.

Due to these shortcomings, we have introduced and thoroughly examined a novel enhancement mode GaN HEMT featuring a MIS [6] and p-GaN structure, referred to as MG-HEMT. The design employs p-GaN layer (forming a schottky contact with the gate of the device) in parallel with HfO_2 layer below the metal gate in the device. By integrating MIS and p-GaN structures side by side within the gate area, we harness the combined advantages such as enhancing gate control, reducing gate leakage, and increasing current capacity. By applying passivation only at the drain side, the capacitance is selectively reduced, minimizing the impact on the overall capacitance and potentially enhancing the device speed and performance. Further, the dielectric layer exhibits a higher bandgap than p-GaN, creating a wider potential barrier for tunneling electrons. This significantly reduces tunneling current through the gate dielectric itself.

Device structure, Calibration and Simulation setup

Figures 1(a) and 1(b) illustrate the schematic representations of the conventional p-GaN HEMT and the proposed MG-HEMT, respectively. In the MG-HEMT configuration, a p-GaN layer is integrated in parallel with an HfO_2 layer, positioned beneath the metal gate within the device structure. Table I shows the parameters of the proposed MG-HEMT structure. For the AlGaN barrier layer, the aluminum content is set at 0.2, and the p-GaN layer has a doping density of $3\times10^{17}/cm^3$ in both the devices.

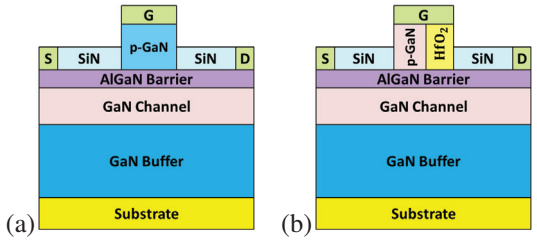


Fig. 1: The schematic diagram of (a) the conventional device and (b) the proposed MG-HEMT.

Calibration of a p-GaN HEMT included parameters like a 2 μm $Al_{0.25}Ga_{0.75}N$ buffer layer, an 85 nm GaN channel layer, a 15 nm $Al_{0.26}Ga_{0.74}N$ barrier layer and a 110 nm thick p-GaN layer [7]. Figure 2 compares experimental and simulated transfer characteristics, showing excellent agreement, validating the precision of our simulation models. Simulation models included the SRH model, GaN polarization model, high-field and low-field mobility models, Fermi–Dirac model,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 14,2024 at 10:03:40 UTC from IEEE Xplore. Restrictions apply.

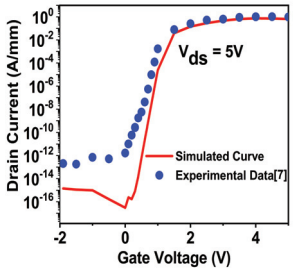


Fig. 2: Calibration of the p-GaN HEMT [7]

TABLE I: The structural parameters of the proposed MG-HEMT

S.No.	Parameter	Value
1	p-GaN layer in parallel with the HfO_2 layer thickness ($T_{Parallel}$)	110 nm
2	AlGaN barrier layer thickness ($T_{AlGaNbarrier}$)	15 nm
3	GaN channel layer thickness ($T_{GaNchannel}$)	85 nm
4	GaN buffer layer thickness ($T_{GaNbuffer}$)	2 μm
5	SiN passivation layer thickness ($T_{SiNpassivation}$)	300 nm

and Toyabe model. Additionally, drift-diffusion transport based on Boltzmann theory accounted for carrier transport in the channel, ensures comprehensive simulation coverage [7].

Results and Discussion

This section investigates the DC and AC characteristics of the device. Figure 3(a) compares the electric field characteristics in the MG-HEMT for varying dielectric constants in the insulator layer. To enhance performance, we have incorporated a high-k dielectric material, known for its superior charge storage capacity compared to low-k materials. This choice reduces the electric field within the device, as high-k materials inherently offer higher capacitance per unit area. Figure 3(b) shows the corresponding leakage current, where the lower electric field decreases the probability of electron tunneling within the device. HfO_2, with its high dielectric constant, exhibits the lowest electric field and gate leakage current, making it the optimized dielectric material for our device.

To optimize device performance, minimizing gate leakage current was crucial for enhancing reliability. We conducted simulations by varying the length of the HfO_2 dielectric material (L_{dt}) to investigate its impact on the device's electrical characteristics. Systematically adjusting L_{dt} from 2.5 μm to 2.8 μm in 0.1 μm increments while keeping other parameters constant revealed that the minimum gate leakage current occurred at 2.7 μm (L_{dt}). This length represents the optimal value for reducing gate leakage and improving device reliability, as depicted in Figure 4(a). Temperature-dependent gate current versus gate voltage (I_g-V_g) curves were analyzed, ranging from 200 K to 400 K, showing that gate leakage current increased with temperature due to rising carrier

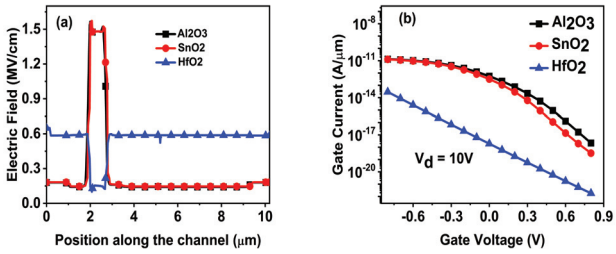


Fig. 3: The comparison of the proposed device with respect to dielectric constant of insulating layer (a) the electric field and (b) the gate leakage current.

concentration and thermal energy, promoting carrier tunneling within the device (shown in Figure 4(b)).

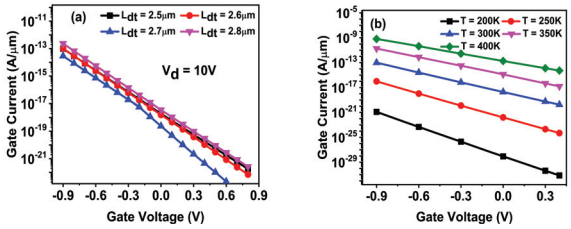


Fig. 4: (a) The gate leakage current for different values of length of the HfO_2 dielectric layer (L_{dt}), and (b) The temperature-dependent I_g-V_g curves of the proposed device from 200K to 400K.

Moreover, we conducted a comparative analysis of the DC and AC characteristics between the conventional p-GaN HEMT and our optimized proposed device. Figure 5(a) displays electric field distributions, showing successful suppression of the electric field in the proposed device's channel due to the HfO_2 dielectric layer [8]. Consequently, our device exhibits lower gate leakage current in Figure 5(b) compared to the conventional device.

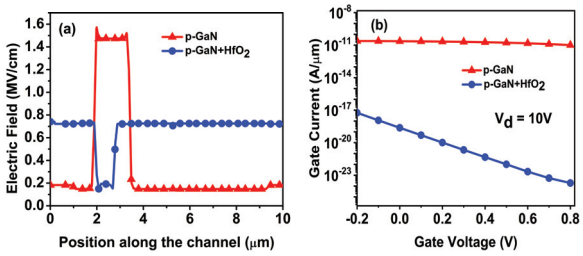


Fig. 5: (a) The electric field and (b) the gate leakage current comparison of the proposed and conventional device.

Figure 6(a) shows the C-V characteristics of our proposed MG-HEMT, highlighting its superior performance over the conventional p-GaN HEMT with reduced parasitic capacitance and lower power consumption. In Figure 6(b), the transconductance comparison reveals that our proposed device outperforms

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 14,2024 at 10:03:40 UTC from IEEE Xplore. Restrictions apply.

TABLE II: Comparison of performance parameters of the proposed device with the reported literature

Reference	Gate leakage current (A/μm)	Maximum drain current (A/μm)	Threshold Voltage (V)
[3]	10^{-15}	0.35	2.5
[4]	10^{-18}	0.8	-15
[5]	10^{-17}	0.7	-5
Conventional p-GaN HEMT	10^{-12}	0.6	0.49
This work	10^{-21}	1	5

the conventional p-GaN HEMT, indicating enhanced current drive capability.

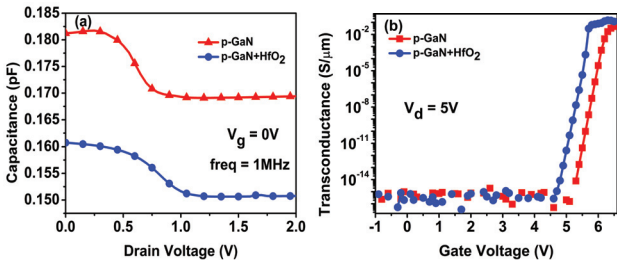


Fig. 6: (a) Gate capacitance and (b) transconductance comparison of the proposed and conventional device.

Figures 7(a) and 7(b) depict the characteristics of I_d - V_d and I_d - V_g, respectively. In Figure 7(a), the proposed MG-HEMT presents a substantially higher maximum drain current, $I_{ds,max}$, reaching 1 A/μm, exceeding the 0.6 A/μm achieved by the conventional p-GaN HEMT. Moreover, the fixed charges can be introduced at the interface or within the HfO_2 layer itself, thereby impacting threshold voltage of the device, as shown in Figure 7(b). Table II depicts the comparison of the proposed MG-HEMT and the already reported devices with respect to their performance parameters.

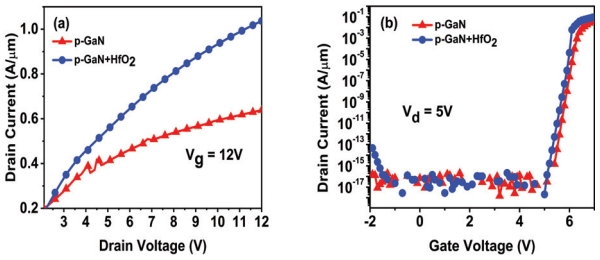


Fig. 7: (a)I_d-V_d and (b)I_d-V_g characteristics comparison of the conventional and proposed device.

CONCLUSION

This paper successfully reports a novel enhancement mode GaN HEMT featuring a MIS and p-GaN structure, referred to as MG-HEMT. The simulation results present a significant decrease of 14% in gate leakage current compared to the already reported device due to low electric field within the proposed device. Also, the proposed device demonstrates 20% and 50% increase in maximum drain current and threshold voltage, respectively, in comparison with the already reported device. These findings indicate that the proposed device shows great potential for power converter application within the field of high-power electronics.

REFERENCES

[1] T. Garg and S. Kale, "A novel p-GaN HEMT with AlInN/AlN/GaN double heterostructure and InAlGaN back-barrier", Microelectronics Reliability, vol. 145, Jun. 2023, doi: 10.1016/j.microrel.2023.114998.

[2] T. Garg and S. Kale, "A novel stepped AlGaN hybrid buffer GaN HEMT for power electronics applications" ,Microelectronics Reliability, vol. 149, Oct. 2023, doi: 10.1016/j.microrel.2023.115232.

[3] H.-C. Chiu et al., "Normally-Off p-GaN Gated AlGaN/GaN MIS-HEMTs with ALD-Grown Al2O3/AlN Composite Gate Insulator," Membranes, vol. 11, no. 10, p. 727, Sep. 2021, doi: 10.3390/membranes11100727.

[4] B. Hult, M. Thorsell, J. -T. Chen and N. Rorsman, "AlGaN/GaN/AlN 'Buffer-Free' High Voltage MISHEMTs with Si-rich and Stoichiometric SiNx First Passivation," 2022 Compound Semiconductor Week (CSW), Ann Arbor, MI, USA, 2022, pp. 1-2, doi: 10.1109/CSW55288.2022.9930464.

[5] M. Whiteside, S. Arulkumaran and G. Ing Ng, "Demonstration of vertically-ordered h-BN/AlGaN/GaN metal-insulator-semiconductor high-electron-mobility transistors on Si substrate", Materials Science and Engineering: B, vol. 270, May 2021, doi: 10.1016/j.mseb.2021.115224.

[6] N. Islam, M. F. P. Mohamed, M. F. A. J. Khan, S. Falina, H. Kawarada, and M. Syamsul, "Reliability, Applications and Challenges of GaN HEMT Technology for Modern Power Devices: A Review," Crystals, vol. 12, no. 11, p. 1581, Nov. 2022, doi: 10.3390/cryst12111581.

[7] Liu, Y., Yu, Q., and Du, J. "Simulation design of a high-breakdown-voltage p-GaN-gate GaN HEMT with a hybrid AlGaN buffer layer for power electronics applications", J Comput Electron, vol. 19, no. 4, pp. 1527–1537, Dec. 2020, doi: 10.1007/s10825-020-01541-2).

[8] A. Wang, L. Zeng and W. Wang "Simulation of Gate Leakage Current of AlGaN/GaN HEMTs: Effects of the Gate Edges and Self-Heating.", ECS Journal of Solid State Science and Technology, vol. 6, no. 11, Jan. 2017, doi: 10.1149/2.0031711jss.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 14,2024 at 10:03:40 UTC from IEEE Xplore. Restrictions apply.

Check for updates

Solutions for Effective Energy Management Through Critical Technologies – Original Research Article

ENERGY EXPLORATION & EXPLOITATION

A novel nature-inspired nutcracker optimizer algorithm for congestion control in power system transmission lines

Energy Exploration & Exploitation
1–36
© The Author(s) 2024
DOI: 10.1177/01445987241253292
journals.sagepub.com/home/eea
Sage

Vivek Kumar[1], R Narendra Rao[2], Ajendra Singh[3], Vineet Shekher[4], Kaushik Paul[4], Pampa Sinha[5], Thamer AH Alghamdi[6,7] and Almoataz Y Abdelaziz[8]

Abstract

In the restructured power system, where uncertainties are common, managing congestion becomes a crucial aspect of power system operation and control. Congestion management aims to alleviate the power system transmission line congestion while meeting the system constraints at minimal cost. This research introduces a generation rescheduling method for congestion management in the electricity market, leveraging an innovative nutcracker optimizer algorithm. The nutcracker optimizer algorithm, inspired by nutcrackers' food accumulation mechanisms, is a recently developed nature-inspired algorithm. The efficacy of this proposed approach is assessed across modified IEEE 30-bus, and IEEE 118-bus test systems, considering the system parameters. The effectiveness of the proposed congestion management with the nutcracker optimizer algorithm is analyzed by comparing its results with those generated by other recent optimization techniques. Results demonstrated that the nutcracker optimizer algorithm surpasses other comparative methods, requiring fewer fitness function evaluations, avoiding local optima, and displaying encouraging convergence traits. Implementing this approach can assist the system

[1]Department of Electrical Engineering, BRCM College of Engineering & Technology, Bhiwani, Haryana, India
[2]Department of Electrical and Electronics Engineering, JNTUA College of Engineering Pulivendula, Pulivendula, India
[3]Department of Electrical Engineering, Delhi Technological University, Delhi, India
[4]Department of Electrical Engineering, BIT Sindri, Dhanbad, India
[5]School of Electrical Engineering, KIIT University, Bhubaneswar, Orissa, India
[6]Electrical Engineering Department, College of Engineering, Al-Baha University, Al-Baha, Saudi Arabia
[7]Wolfson Centre for Magnetics, School of Engineering, Cardiff University, Cardiff, UK
[8]Faculty of Engineering and Technology, Future University in Egypt, Cairo, Egypt

Corresponding author:
Kaushik Paul, Department of Electrical Engineering, BIT Sindri, Dhanbad 828123, India.
Email: kaushik.ee@bitsindri.ac.in

Thamer AH Alghamdi, Wolfson Centre for Magnetics, School of Engineering, Cardiff University, Cardiff, CF24 3AA, UK.
Email: Alghamdit1@cardiff.ac.uk

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).

operators in swiftly addressing contingencies, ensuring secure and reliable power system operation within a deregulated environment.

Keywords
Energy, power flow, evolutionary algorithm, power system, optimization, rescheduling

Introduction

The ability of transmission lines to carry electrical power is limited by various factors, including voltage, heat, and stability. When any of these limits is reached, the power system is considered congested. Congestion can have significant implications on the power dispatch, potentially impacting the overall performance of the system. Managing power system congestion is a critical task undertaken by independent system operators (ISOs) and grid operators (Bouhouras et al., 2024). They play a crucial role in coordinating and overseeing the functioning of the power market, ensuring efficient power flow, and maintaining system stability. With the increasing demand for electricity, aging infrastructure, and limitations on new transmission line construction, congestion has become a prevalent issue in power systems worldwide (Alavijeh et al., 2024).

Several factors contribute to congestion. These include an upsurge in power demand beyond the infrastructure's capacity, unexpected generation outages, limitations on building new transmission lines due to environmental or logistical constraints, unscheduled power flow, transmission line tripping, or equipment failures. The resulting congestion can lead to reduced reliability, increased energy costs, and potential disruptions in power supply to consumers (Dehnavi et al., 2024). To address congestion effectively, various congestion management (CM) strategies are employed. These strategies aim to alleviate constraints, optimize power flow, and maintain system stability. Some common approaches include reactive power support, generation rescheduling, demand response programs, transmission line reconfiguration, and dispatching of power resources (Ansari et al., 2020; Shaikh et al., 2020; Zhou et al., 2023). These methods help balance the power system and mitigate congestion-related issues, ensuring reliable and uninterrupted electricity supply to consumers.

Literature survey

In recent years, advancements in technology, such as smart grid technologies, advanced metering infrastructure, and real-time monitoring systems, have provided enhanced capabilities for CM. These tools enable grid operators to gather precise data, monitor system conditions, and implement timely control actions to alleviate congestion and optimize power flow. Zaidan and Toos proposed an emergency CM method for power systems using a static synchronous series compensator (SSSC). The authors focused on alleviating congestion during emergency situations and introduced the SSSC as a device to enhance the power transfer capability of transmission lines (Majeed Rashid Zaidan, 2022). The research focuses on optimizing transmission line parameters using an enhanced moth flame optimization technique, demonstrating its superiority over other methods through simulations on various test cases. Results indicate that the proposed method achieves rapid convergence to optimal values for capacitance and inductance, showcasing its effectiveness and competence in managing transmission and distribution networks (Shaikh et al., 2022c). Wang et al. introduced the concept of virtual storage plants (VSPs) for frequency regulation and CM in power systems. They proposed a framework where distributed energy resources are aggregated into VSPs to provide

ancillary services for frequency regulation and CM. The study evaluates the performance of the VSP and its potential for improving the stability and efficiency of power systems (Wang et al., 2022). Managing the transmission line flows with the influence of optimization algorithm for transmission line parameter optimization can be found in (Shaikh et al., 2021a, 2022a, 2022b, 2023). Mahajan et al. provided a review of CM in deregulated power systems. The review associated with the CM summarized the existing literature on CM techniques, including market-based approaches, generation rescheduling, load shedding, and demand response programs. The review also provided insights into the challenges and opportunities in managing congestion in deregulated power systems (Mahajan et al., 2022). Hobbie et al. investigated the effect of model parametrization of the transmission channels during the congested environment of the power system. The authors compare different existing models used for CM in electricity networks and analyze the influence of model parameters on CM (Hobbie et al., 2022). Zakaryaseraji and Ghasemi-Marzbali evaluated CM in power systems considering demand response programs and distributed generation (DG). They incorporated demand response and DG as flexible resources for managing congestion (Zakaryaseraji and Ghasemi-Marzbali, 2022). Shaikh et al. conducted their research to select the optimal transmission-line parameters that would assist efficient power flows (Shaikh et al., 2021b). Roustaei et al. formulated a voltage security-constrained CM methodology for improving the controlling of transmission systems. They addressed the challenges of voltage security and CM and proposed an approach that optimizes generation rescheduling to alleviate congestion while considering voltage security constraints (Roustaei et al., 2022). In another research, Mishra et al. provided a review of the utilization of AI-based methods in alleviating the congestion of transmission channels considering the electricity market constraints. The review mainly highlighted the application of artificial intelligence techniques, such as neural networks, fuzzy logic, and evolutionary algorithms, for CM in restructured power systems (Mishra et al., 2022). Paul formulated an enhanced framework of a grey wolf optimization algorithm that controls the congestion considering the impact of renewable sources. The CM focused on optimizing the operation of transmission lines to alleviate congestion and incorporates the impact of solar photovoltaic systems in the optimization process (Paul, 2022a). Dehnavi et al. proposed a novel CM method through power system partitioning. In the CM approach, they considered the power system into multiple partitions and applied a decentralized control strategy to manage congestion (Dehnavi et al., 2022). Sarwar et al. developed an effective transmission CM approach for contemporary power system operations using optimal DG capacity and hybrid swarm optimization. The authors focused on determining the optimal capacity of DG units at specific locations in the power system to mitigate congestion (Sarwar et al., 2022).

In the CM, generator rescheduling (GR) plays a crucial role by optimizing the generation dispatch to alleviate congestion. It allows for the efficient utilization of existing transmission infrastructure, reducing the need for costly grid expansion. By strategically adjusting generator schedules, congestion can be alleviated, ensuring reliable and secure power system operation. GR also enables the integration of renewable energy sources and promotes the economic dispatch of generation resources, leading to enhanced system efficiency and cost-effectiveness. Subramaniyan and Gomathi presented a soft computing-based analysis of CM in transmission systems with the application of generator rescheduling. The CM approach explored the application of soft computing techniques, such as fuzzy logic and genetic algorithms, for managing congestion (Subramaniyan and Gomathi, 2023). Agrawal et al. proposed a cascaded deep neural network-based approach for customer participation in CM, considering renewable energy sources and GR in deregulated power markets. Their research focused on optimizing the participation of customers with distributed energy resources to alleviate congestion (Agrawal et al., 2022). Chakravarthi et al. investigated real-time CM using generation redispatch. They designed controllers for efficient

generation rescheduling to mitigate congestion in power systems (Chakravarthi et al., 2022). Srivastava and Yadav proposed a rescheduling-based CM approach using a hybrid algorithm that combines lion and moth search models. The authors optimized generator schedules to alleviate congestion and enhance the operational efficiency of power systems (Srivastava and Yadav, 2021). Ogunwole and Krishnamurthy developed a transmission CM approach using generator sensitivity factors and particle swarm optimization. The authors focus on optimizing active and reactive power rescheduling to alleviate congestion and enhance system performance (Ogunwole and Krishnamurthy, 2022). Prajapati et al. proposed a GR-based CM approach considering the demand response characteristic for power systems and uncertainty of the renewable resources (Thiruvel et al., 2023) and presented a modern power system operations approach for effective transmission CM using optimal distributed generation (DG) capacity and GR. The authors determined the optimal capacity of DG units and real power deliveries by the generators to alleviate congestion (Thiruvel et al., 2023). Balaraman and Kamaraj managed congestion for a deregulated electricity market with the application of GR methodology. They implemented particle swarm optimization (PSO) to alleviate overloads and minimize deviations from the initial market settlement. The PSO application for CM addressed system constraints security constraints, using a penalty-based approach (Kamaraj, 2011). Saravanan and Anbalagan proposed an intelligent hybrid technique for optimal GR in deregulated power markets. They combined genetic algorithms (GA) and PSO to optimize generator schedules and alleviate congestion (Saravanan and Anbalagan, 2022). Sharma and Walde proposed a novel optimization technique based on swarm intelligence for CM in transmission lines. They applied a hybrid algorithm combining PSO and GSA to optimize power flow and alleviate congestion (Sharma and Walde, 2022). Haq et al. proposed a game-theoretic approach using plug-in electric vehicles (PEVs) and GR for real-time CM. Their work investigated the participation of PEVs and GR in alleviating congestion by adjusting their charging and discharging patterns (Haq et al., 2022). Verma and Mukherjee implemented real power GR rescheduling using a novel ant lion optimizer for CM. The CM approach aimed at optimizing generator schedules to alleviate congestion and enhance power system performance (Verma and Mukherjee, 2016b). Paul et al. reduced rescheduling costs and efficiently alleviated line overloads with minimal deviations in active power generation. The developed CM strategy involved prioritizing generators based on their sensitivity values. The GSA is then employed to minimize rescheduling costs while simultaneously reducing total active power output and system losses (Paul et al., 2019).

The emergence of competitive energy markets has driven the advancement of optimization algorithms that effectively tackle mathematical problems related to power system challenges, resulting in precise and improved outcomes. The classical optimization methods have been widely used and effective in certain scenarios but have some limitations that make metaheuristic optimization methods more preferable in many cases. Classical optimization methods have limitations such as getting entombed in local optima, sensitivity to initial conditions, lack of flexibility, high computational cost, and difficulty in handling uncertainties and noisy environments. In contrast, metaheuristic algorithms overcome these limitations by exploring the entire solution space, being less sensitive to initial conditions, offering flexibility in problem domains, providing computational efficiency, and accommodating uncertainties and noisy environments through robust and adaptive mechanisms. The metaheuristic algorithms strike a balance between exploration (diversification) and exploitation (intensification) of the search space. They avoid getting stuck in local optima by exploring different regions while also exploiting promising areas. This ability to balance exploration and exploitation enhances their capability to find better solutions. Various approaches such as grey wolf optimization (Paul, 2022b), Fennec Fox optimization (Verma and Mukherjee, 2016a), flower pollination algorithm (FPA; Gupta et al., 2023), cuckoo search algorithm (CSA; Wangunyu et al., 2022), and an enhanced version of crow search algorithm (CSA) (Paul et al., 2022) have been implemented to oversee the efficient operation of power

systems. The selection and adjustment of control parameters associated with these meta-heuristic techniques greatly influence the achieved solution. This has motivated the development of an efficient CM strategy with the application of the evolutionary algorithm that will deliver optimal outcomes for the optimization problem.

The novelty can be stated as the minimization of the congestion cost/rescheduling cost always creates a scope for implementation of new efficient optimization approaches that would deliver appreciable outcomes w.r.t previously adopted approaches. The integral novelty of this paper lies in the mathematical model formulation of the congestion cost minimization with the incorporation of the new nutcracker optimization algorithm (NOA).

In this research work, NOA which is a new metaheuristic optimization technique has been considered for the CM problem. The NOA has been designed by Basset et al. in the year 2023 (Abdel-Basset et al., 2023). The framework of NOA is inspired by the intelligent behaviors of nutcrackers to search for and store food. It draws inspiration from the hunting behavior of antlions. According to Abdel-Basset et al. (2023), the NOA has been validated on standard mathematical functions (CEC-2014, CEC-2017, and CEC-2020) and classical engineering problems. The findings presented by Abdel-Basset et al. (2023) demonstrated the advantages of the NOA in terms of improved exploration capabilities.

Motivation

The primary motivation behind this endeavor is to devise a proficient approach for CM aimed at enhancing the efficiency of power system operations. The competitive dynamics and trading behavior within the electricity market have heightened the vulnerability of the power network to congestion, prompting researchers in the field to investigate various CM strategies for their applicability and effectiveness. Given the extensive scale of the power network, any implemented CM strategy must possess qualities of robustness, effectiveness, and cost-efficiency, and must not disrupt the network's existing topology. Thus, in light of these considerations, the implementation of the generators rescheduling (GR) approach emerges as a fitting solution for CM. Furthermore, in power system operations, it's imperative to mitigate congestion at minimal cost, necessitating the utilization of efficient optimization techniques that yield optimal cost solutions. While many optimization techniques exhibit commendable performance, they often become ensnared in local optima, unable to navigate the search space effectively to identify the optimal solution. Consequently, there's a strong impetus to enhance the performance of optimization techniques by refining their operational stages. This drive has spurred advancements in the efficiency of exploration and exploitation stages within the NOA.

Research contribution

- Development of a CM model to minimize the congestion cost based on the optimal rescheduling of the generators' real power.
- Implementation of NOA to achieve the optimal power rescheduling patterns for generators for congestion cost minimization.
- Application of the CM strategies on power stream standard test system; IEEE 30-bus and IEEE 118-bus to validate the efficiency of the proposed CM model.
- Establish comparative performance analysis based on the congestion cost, bus voltage profile, and system losses achieved with NOA with the recent optimization techniques.

Problem formulation

The primary goal of CM is to reduce the expenses caused by congestion while adhering to the limitations of the network. This study addresses the CM problem by adjusting the actual power output of generators, either increasing or decreasing it. However, modifying the real power output incurs costs that are decided based on the price offers provided by the generating companies. This can be represented as.

$$\min CC = \sum_{n=1}^{N_g} (C_{inc} \cdot \Delta P_g + C_{dec} \cdot \Delta P_g) \, \$ / \mathrm{h} \tag{1}$$

The equation involves various variables: CC represents the aggregated cost of adjusting real power in \$/h, C_{inc} and C_{dec} denotes the incremental and decremental price bids, respectively, ΔP_g is the adjusted power by the generators. The constraints considered for this optimization problem are as follows:

Inequality constraints

These constraints are expressed through the following equations:

$$P_{Gn}^{\min} \le P_{Gn} \le P_{Gn}^{\max}; \quad \forall n \in N_g \tag{2}$$

$$Q_{Gn}^{\min} \le Q_{Gn} \le Q_{Gn}^{\max}; \quad \forall n \in N_g \tag{3}$$

$$(P_{Gn} - P_{Gn}^{\min}) = \Delta P_{Gn}^{\min} \le \Delta P_{Gn} \le \Delta P_{Gn}^{\max} = (P_{Gn}^{\max} - P_{Gn}); \quad \forall n \in N_g \tag{4}$$

$$V_n^{\min} \le V_n \le V_n^{\max} \tag{5}$$

$$P_{ij} \le P_{ij}^{\max} \tag{6}$$

Equality constraints

$$P_{Gn} - P_{Dn} = \sum_j |V_j||V_k||Y_{kj}| \cos(\delta_k - \delta_j - \theta_{kj}); \quad j = 1, 2, 3, \ldots, N_b \tag{7}$$

$$Q_{Gn} - Q_{Dn} = \sum_j |V_j||V_k||Y_{kj}| \sin(\delta_k - \delta_j - \theta_{kj}); \quad j = 1, 2, 3, \ldots, N_b \tag{8}$$

$$P_{Gn} = P_{Gn}^{C} + \Delta P_{Gn}^{+} - \Delta P_{Gn}^{-}; \quad k = 1, 2, 3, \ldots, N_g \tag{9}$$

$$P_{Dj} = P_{Dj}^{C}; \quad j = 1, 2, 3, \ldots, N_d \tag{10}$$

Here, in equations (7) and (8), P_{Gk} and Q_{Gk} represent the active and reactive powers generated at the kth bus, respectively. Similarly, P_{Dk} and Q_{Dk} denote the active and reactive powers at the kth bus,

respectively. In equations (7) and (8), V_j and V_k indicate the voltages at buses, while δ_k and δ_j are the bus voltage angles. The parameter θ_{kj} represents the admittance angle. The count of the generators, buses and loads are given by N_b, N_g, and N_d, respectively. P_{Gk}^{C} is the generated power and P_{Dj}^{C} is the consumed by the load. It should be noted that equations (7) and (8) outline the balances of active and reactive power at individual nodes, whereas equations (9) and (10) portray the resulting power in relation to the market clearing price.

Nutcracker optimizer algorithm

This section presents a bio-inspired optimization algorithm (NOA) that draws inspiration from the intelligent behaviors of nutcrackers. The NOA has been designed by Basset et al. in the year 2023 (Abdel-Basset et al., 2023). The nutcracker's behavior can be segregated and portrayed into two sections. The first part involves collecting and storing pine seeds as food. The second part involves searching for and retrieving the stored seeds.

Foraging stage: Exploration phase I

In the starting position, each nutcracker examines the cone holding seeds; if good seeds are found, they are taken to the storage area for burial. If no good seeds are found, the nutcracker seeks another cone among pine/trees. The mathematical representation of this behavior representing the position update can be stated as

$$\overrightarrow{Z}_i^{t+1} = \begin{cases} Z_{i,j}^{t}; & \text{if } \tau_1 < \tau_2 \\ Z_{m,j}^{t} + \psi(Z_{A,j}^{t} - Z_{B,j}^{t}) + \xi(r^2 U_j - L_j); & \text{if } t \le T_{\max}/2.0 \\ Z_{C,j}^{t} + \xi(Z_{A,j}^{t} - Z_{B,j}^{t}) + \xi(r^2 U_j - L_j); & \text{otherwise} \end{cases} \tag{11}$$

$\overrightarrow{Z}^{t+1}$ denotes the current position of the ith nutcracker, and $Z_{i,j}^{t}$ represents its jth position. Vectors U_j and L_j define the upper and lower bounds of the jth dimension in the optimization problem, respectively. ψ is a randomly generated number using the levy flight. $\overrightarrow{Z}_{best,j}^{t}$ signifies the jth dimension of the best solution. τ_1, τ_2, and r ranges [0, 1]. $Z_{m,j}^{t}$ is the mean of all solutions. In equation (12), ξ is generated based on the normal distribution (τ_4), levy-flight τ_5, and τ_3 [0, 1].

Top of form

$$\xi = \begin{cases} \tau_3 \text{ if } r_1 < r_1 \\ \tau_4 \text{ if } r_2 < r_3 \\ \tau_5 \text{ if } r_1 < r_3 \end{cases} \tag{12}$$

In this scenario, r_1, r_2, and r_3 are arbitrary real numbers between 0 and 1. Equation (11) has been created to study top-notch food options. The initial part of equation (11) shows how likely it is for nutcrackers to find good seeds on their first try, suggesting they won't change the original dimension in the solution Z_i^t. Equation's (11) second stage helps nutcrackers explore the search space globally, boosting NOA's search power. In the third stage, nutcrackers explore positions near a randomly chosen solution while keeping the chance to move globally with a δ probability. A and B offer data on new food spots, while τ_1 and τ_2 regulate NOA's exploration. The algorithm switches between local and global searches using diverse τ_1 and τ_2 values. μ guides nutcrackers with directions and step sizes for exploring new positions.

Storage stage: Exploitation phase 1

Initially, nutcrackers traverse with the gathered foodstuff in the prior exploration phase, called phase 1, to interim stowing in the storage area. This phase is termed "exploitation phase 1" as it involves both exploiting and storing pine seed crops. This behavior can be mathematically represented as:

$$\overrightarrow{Z}_i^{t+1(new)} = \begin{cases} \overrightarrow{Z}_i^t + \xi \cdot (\overrightarrow{Z}_{best}^t - \overrightarrow{Z}_i^t) \cdot |\gamma| + r_1 \cdot (\overrightarrow{Z}_A^t - \overrightarrow{Z}_B^t) & \text{if } \tau_1 < \tau_2 \\ \overrightarrow{Z}_{best}^t + \xi \cdot (\overrightarrow{Z}_A^t - \overrightarrow{Z}_B^t) & \text{if } \tau_1 < \tau_3 \\ \overrightarrow{Z}_{best}^t \cdot l & \text{otherwise} \end{cases} \tag{13}$$

In the current iteration, the position of a new item in the storage area of the nutcrackers is represented by a vector $\overrightarrow{Z}_i^{t+1(new)}$. $\overrightarrow{Z}_{best}^t$ represents the jth dimension of the best solution obtained so far. The value of γ is determined by a Lévy flight, while τ_3 is a random number between 0 and 1. The factor l gradually decreases from 1 to 0, aiding in broadening NOA's exploitation behavior, this diversity in approach speeds up convergence and avoids getting stuck in local minimums that may occur when exploring in just one direction.

Throughout the optimization process, a consistent equilibrium between exploration and exploitation operators is upheld by applying the following formula to regulate the interaction between the foraging stage and the cache:

$$\overrightarrow{Z}_i^{t+1} = \begin{cases} \text{Equation (11), if } \varphi > P_{a1} \\ \text{Equation (13), otherwise} \end{cases} \tag{14}$$

In this stage, a random number ϕ is generated between zero and one, and the probability value P_{a1} gradually decreases from one to zero. The generated solutions are directed towards a specific region to cover as many solutions as possible.

Cache-search and recovery strategy

- Cache-search stage: Exploration phase 2: Nutcrackers employ spatial memory tactics to uncover their caches, where multiple objects might signify a single cache. Each cache corresponds to just two markers or objects, known as reference points (RPs). Within the NOA, two RPs are established for each population's cache/nutcracker using the following matrix:

$$RPs = \begin{bmatrix} RP_{1,1}^t & RP_{1,2}^t \\ \vdots & \vdots \\ RP_{i,1}^t & RP_{i,2}^t \\ \vdots & \vdots \\ RP_{n,1}^t & RP_{n,2}^t \\ \vdots & \vdots \end{bmatrix} \tag{15}$$

In this equation, $RP_{i,1}^t$ and $RP_{i,1}^t$ stand for the RPs linked to the ith nutcracker's cache position $\overrightarrow{Z}_i^t$ at generation t.

Nutcrackers find caches accurately, but about 20% of attempts fail initially. If they cannot locate the food with the first RP, they try the second RP. Failing that, they turn to the third reference. Two equations boost nutcrackers' exploration for caches. The first *RP* updates the position to uncover nearby caches. Its formula is

$$\overrightarrow{RP}_{i,k}^{t} = \overrightarrow{Z}_{i}^{t} + \alpha \cdot \cos(\theta) \cdot ((\overrightarrow{Z}_{A}^{t} - \overrightarrow{Z}_{B}^{t})), k = 1 \tag{16}$$

The second RP adjusts the existing solution to assist nutcrackers explore diverse areas. Its calculation follows a specific formula:

$$\overrightarrow{RP}_{i,k}^{t} = \overrightarrow{Z}_{i}^{t} + \alpha \cdot \cos(\theta) \cdot ((\overrightarrow{U} - \overrightarrow{L}) \cdot \tau_3 + \overrightarrow{L}) \cdot \overrightarrow{U}_2; k = 2 \tag{17}$$

$$\overrightarrow{U} = \begin{cases} 1; \overrightarrow{r_2} < P_{rp} \\ 0; \text{ otherwise} \end{cases} \tag{18}$$

Equations (16) and (17) create two RPs for each nutcracker, aiding in cache recovery. Close RPs make retrieval easier; initially, nutcrackers lack experience but learn over time. In the equations, α trains nutcrackers, preventing RPs from converging too soon. θ, randomly picked from 0 to π. When $\theta = \pi/2$, $\overrightarrow{RP}_{i,k}^{t}$ matches $\overrightarrow{Z}_{i}^{t}$ which portrays their overlap. To handle this, equations (16) and (17) can be combined and represented as

$$\overrightarrow{RP}_{i,1}^{t} = \begin{cases} \overrightarrow{Z}_{i}^{t} + \alpha \cdot \cos(\theta) \cdot ((\overrightarrow{Z}_{A}^{t} - \overrightarrow{Z}_{B}^{t})) + \alpha \cdot RP; & \text{if } \theta = \pi / 2 \\ \overrightarrow{Z}_{i}^{t} + \alpha \cdot \cos(\theta) \cdot ((\overrightarrow{Z}_{A}^{t} - \overrightarrow{Z}_{B}^{t})); & \text{otherwise} \end{cases} \tag{19}$$

$$\overrightarrow{RP}_{i,2}^{t} = \begin{cases} \overrightarrow{Z}_{i}^{t} + (\alpha \cdot \cos(\theta) \cdot ((\overrightarrow{U} - \overrightarrow{L}) \cdot \tau_3 + \overrightarrow{L}) + \alpha \cdot RP) \cdot \overrightarrow{U}_2; & \text{if } \theta = \pi / 2 \\ \overrightarrow{Z}_{i}^{t} + \alpha \cdot \cos(\theta) \cdot ((\overrightarrow{U} - \overrightarrow{L}) \cdot \tau_3 + \overrightarrow{L}) \cdot \overrightarrow{U}_2; & \text{otherwise} \end{cases} \tag{20}$$

$\overrightarrow{RP}_{i,1}^{t}$ refers to the element in the *i*th row and first column of the matrix in equation (15), while $\overrightarrow{RP}_{i,2}^{t}$ is in the *i*th row and second column of the same matrix. Each row corresponds to a cache/nutcracker, each column to an *RP*. *RP* means a random position. To avoid *RPs* converging too soon, the third term in equations (19) and (20) ($\theta = \pi/2$) is added. α ensures regular convergence in NOA, improving RP selection in later generations. Calculate α with this equation:

$$\alpha = \begin{cases} \left(1 - \frac{t}{T_{\max}}\right)^{2 \cdot \frac{t}{T_{\max}}}; & \text{if } r_1 > r_2 \\ \left(\frac{t}{T_{\max}}\right)^{\frac{2}{t}}; & \text{otherwise} \end{cases} \tag{21}$$

In this case, t is the current generation, and $T_{\max}$ is the maximum number of generations. The first stage in equation (21) decreases with each round, speeding up convergence. Meanwhile, the second stage rises gradually to avoid trapping the algorithm in local lows caused by the first stage.

During optimization, nutcrackers gain expertise in *RP* selection. They adjust their storage based on these RPs. In NOA, all nutcrackers explore promising regions using an exploration mechanism. Progressively, the algorithm avoids local lows by dynamically exploring and exploiting areas

around caches with suitable *RPs*. The nutcracker's updated position follows this equation:

$$\overrightarrow{Z}_i^{t+1} = \begin{cases} \overrightarrow{Z}_i^t; & \text{if } f(\overrightarrow{Z}_i^t) \text{lt} (\overrightarrow{RP}_{i,1}^t) \\ \overrightarrow{RP}_{i,1}^t; & \text{otherwise} \end{cases} \tag{22}$$

The equation demonstrates the new cache $\overrightarrow{Z}_i^{t+1}$ of the *i*th nutcracker in iteration $(t+1)$, where $\overrightarrow{Z}_i^t$ represents the current cache, and $\overrightarrow{RP}_{i,1}^t$ is the initial link associated with the current cache in iteration *t*. Equation (22) steers NOA toward exploring and exploiting areas near $\overrightarrow{RP}_{i,1}^t$.

- *Recovery stage: Exploitation phase 2:* The main scenario: a nutcracker recalls the cache with the first RP. Figure 8 shows two options: food present or absent. Mathematically, it's expressed as follows:

$$\overrightarrow{Z}_{i,j}^{t+1} = \begin{cases} \overrightarrow{Z}_{i,j}^t; & \text{if } \tau_3 < \tau_4 \\ \overrightarrow{Z}_{i,j}^t + r_1 \cdot (\overrightarrow{Z}_{best,j}^t - \overrightarrow{Z}_{i,j}^t) + r_2 (\overrightarrow{RP}_{i,j}^t - \overrightarrow{Z}_{i,j}^t); & \text{otherwise} \end{cases} \tag{23}$$

Let $\overrightarrow{Z}_{i,j}^{t+1}$ be the updated cache, $\overrightarrow{Z}_{i,j}^t$ is the current dimension, and $\overrightarrow{Z}_{best,j}^t$ shows the best cache found so far in iteration *t*. $\overrightarrow{RP}_{i,j}^t$ is the random parameter of the *i*th nutcracker's current cache in iteration *t*. r_1, r_2, τ_3, and τ_4 ranges between [0, 1].

Equation's (23) first condition assumes food is present, passing certain dimensions to the consecutive generation. The second condition in equation (23) assumes no food, signaling an unpromising solution; an escape mechanism avoids local lows. Food loss in the cache could be due to theft or natural elements. The first part of equation (23) helps the nutcracker exploit promising areas locally. The second part allows exploration globally in the search space.

If the nutcracker cannot find food with the first *RP*, it tries the second one. Nutcrackers store multiple *RPs* in early storage, but sometimes the first *RP* fails. Some rely on changing landmarks when autumn becomes winter. Equation (24) updates nutcracker memory with the second *RP*.

$$\overrightarrow{Z}_i^{t+1} = \begin{cases} \overrightarrow{Z}_i^t; & \text{if } f(\overrightarrow{Z}_i^t) \text{lt} (\overrightarrow{RP}_{i,2}^t) \\ \overrightarrow{RP}_{i,2}^t; & \text{otherwise} \end{cases} \tag{24}$$

In this setting, $\overrightarrow{Z}_i^t$ is the current cache, and $\overrightarrow{RP}_{i,2}^t$ is the second *RP* linked to the *i*th nutcracker. Equation (24) enables NOA to observe new locations within the vicinity of the second *RP* and perform exploitation within the most prominent areas. In NOA, finding the cache with the second *RP* updates equation (13) based on this information.

$$\overrightarrow{Z}_{i,j}^{t+1} = \begin{cases} \overrightarrow{Z}_{i,j}^t; & \text{if } \tau_5 < \tau_6 \\ \overrightarrow{Z}_{i,j}^t + r_1 \cdot (\overrightarrow{Z}_{best,j}^t - \overrightarrow{Z}_{i,j}^t) + r_2 (\overrightarrow{RP}_{i,2}^t - \overrightarrow{Z}_{Cj}^t); & \text{otherwise} \end{cases} \tag{25}$$

In the given equation, r_1, r_2, τ_5, and τ_6 are random numbers from 0 to 1. Equation's (25) first part intensifies local search in optimal areas. The second part explores new spaces for a better global

search.

$$\overrightarrow{Z}_i^{t+1} = \begin{cases} \text{Equation (13), if } \tau_7 < \tau_8 \\ \text{Equation (15), otherwise} \end{cases} \tag{26}$$

In this equation, τ_7 and τ_8 are random numbers from 0 to 1. First case: nutcracker recalls the store; second: cannot remember. Equation balances exploration between the two RPs:

$$\overrightarrow{Z}_i^{t+1} = \begin{cases} \text{Equation (12)}, & \text{if } f(\text{equation (12)}) < f(\text{equation (14)}) \\ \text{Equation (14)}, & \text{otherwise} \end{cases} \tag{27}$$

To sustain a steadiness between exploration and exploitation, the transition between the cache-search phase and the recovery phase is executed through the utilization of the subsequent formula.

$$\overrightarrow{Z}_i^{t+1} = \begin{cases} \text{Equation (16), if } \phi < P_{a2} \\ \text{Equation (17), otherwise} \end{cases} \tag{28}$$

In this equation, φ ranges from 0 to 1, and P_{a2} is set at 0.2 based on several trial runs, and the best result is obtained from stand benchmark functions.

Performance analysis of NOA

The NOA's performance has been evaluated across 23 standard benchmark functions, including seven unimodal and 13 multimodal functions. The details of these functions can be found by Dolatabadi et al. (2020). Comparisons with GTO, AHA, and the original NOA are conducted. In this study, search agents and maximum iterations are set at 40 and 500, respectively, with findings gathered from 30 separate runs. Results show that NOA exhibits significant performance during both the exploration and exploitation stages. Table 1 provides the average (Avg), standard deviation (SD), median (Med), and worst (worst of the best-so-far solution) values obtained for GTO, AHA, and NOA the original MRFO in the final repetition. The test results with 23 benchmark functions are given in Table 1. Convergence profiles of the 23 benchmark functions using NOA, AHA, and GTO are depicted in Figure 1, while Figure 2 illustrates the flowchart for NOA in addressing the proposed CM problem.

Computational complexity of NOA

The evaluation of algorithm complexity in dimensions 20, 40, and 50 is conducted according to the guidelines outlined by Liang et al. (2013) and the results are illustrated in Table 2. The runtime is determined through the assessment of the CEC 2014 benchmark function FCEC18. The determination of T_0 is accomplished resulting in a computed execution time of $T_0 = 0.7539$ s. T_1 signifies the duration required for executing 2.00×10^5 evaluations of the benchmark function FCEC18 individually across dimensions of 20, 40, and 50. Meanwhile, T_2 denotes the time taken for executing an algorithm encompassing 2.00×10^5 evaluations of FCEC18 in 20, 40, and 50 dimensions.

Results

The proposed NOA is tested on three different power systems: the modified IEEE 30-bus (Sharma et al., 2018) and IEEE 118-bus test systems (Dash and Mohanty, 2015). The necessary data for these systems is obtained from Paul (2022b) and Paul et al. (2022). In addition to applying the

Table 1. Performance of NOA, GTO, and AHA on 23 benchmark functions.

Functions		NOA	GTO	AHA
F1	Avg	0	1.82×10^{-43}	2.77×10^{-27}
	SD	0	1.9×10^{-43}	8.68×10^{-27}
	Med	0	1.56×10^{-43}	2.27×10^{-29}
	Worst	0	7.9×10^{-43}	4.59×10^{-26}
F2	Avg	0	7.61×10^{-26}	1.27×10^{-19}
	SD	0	7.12×10^{-26}	3.05×10^{-20}
	Med	0	6.19×10^{-26}	7.67×10^{-23}
	Worst	0	2.72×10^{-25}	7.94×10^{-20}
F3	Avg	0	0.186	4.58×10^{-8}
	SD	0	0.133	0.000000128
	Med	0	0.131	3.95×10^{-12}
	Worst	0	0.131	0.000000407
F4	Avg	0	0.000000211	2.54×10^{-9}
	SD	0	8.88×10^{-8}	5.46×10^{-9}
	Med	0	0.00000019	4.81×10^{-11}
	Worst	0	0.000000339	0.000000168
F5	Avg	931.2192	3.2	7.118
	SD	2396.3352	2.2	0.25018
	Med	135.1696	2.78	7.2289
	Worst	7710.4457	6.68	7.3661
F6	Avg	0	0	0.31179
	SD	0	0	0.11916
	Med	0	0	0.29383
	Worst	0	0	0.48389
F7	Avg	0.0000157	0.0028	0.0007298
	SD	0.0000141	0.00103	0.0005702
	Med	0.000012	0.0026	0.0006444
	Worst	0.0000531	0.00478	0.0020895
F8	Avg	−2665.515	−4170	−2313.3951
	SD	223.8271	50	248.3506
	Med	−2634.759	−4190	−2197.3255
	Worst	−2400.079	−4070	−2136.702
F9	Avg	0	0	1.5968
	SD	0	0	5.0495
	Med	0	0	6.65×10^{-10}
	Worst	0	0	15.9679
F10	Avg	8.79×10^{-17}	5.55×10^{-16}	3.5×10^{-15}
	SD	0	0	4.15×10^{-14}
	Med	9.99×10^{-16}	5.55×10^{-15}	8.07×10^{-15}
	Worst	9.09×10^{-16}	5.55×10^{-15}	2×10^{-13}
F11	Avg	0	0.00074	0.090781
	SD	0	0.00234	0.19771
	Med	0	0	0.00000098
	Worst	0	0.0074	0.61623
F12	Avg	0.10412	4.71×10^{-32}	0.070837
	SD	0.43889	2.26×10^{-48}	0.020720

(continued)

Table 1. Continued.

Functions		NOA	GTO	AHA
	Med	0.0000829	5.81×10^{-33}	0.086773
	Worst	2.148	5.81×10^{-33}	0.006228
F13	Avg	0.010807	1.35×10^{-32}	0.22513
	SD	0.026982	3.90×10^{-49}	0.08226
	Med	0.0072202	2.46×10^{-33}	0.35616
	Worst	0.061285	2.46×10^{-33}	0.43123
F14	Avg	0.998	0.998	1.5935
	SD	0	0	0.95823
	Med	0.998	0.998	0.99809
	Worst	0.998	0.998	2.9821
F15	Avg	0.0003828	0.000749	0.0008991
	SD	0.000238	0.000191	0.000331
	Med	0.0003075	0.000718	0.0007456
	Worst	0.0010602	0.00122	0.0014997
F16	Avg	−1.03161	−1.03	−1.0316
	SD	0	0	0.000016
	Med	−2.14272	−1.03	−2.1427
	Worst	−2.14272	−1.03	−2.1427
F17	Avg	0.40890	0.398	0.39898
	SD	0	0	0.0011636
	Med	0.40890	0.409	0.40946
	Worst	0.40890	0.398	0.40126
F18	Avg	2	2	2
	SD	0	2.96×10^{-16}	0.0000396
	Med	2	2	2
	Worst	2	2	2.0001
F19	Avg	−4.9739	−4.97	−4.9651
	SD	0	9.47×10^{-17}	0.0007109
	Med	−4.9719	−4.97	−4.9656
	Worst	−4.9719	−4.97	−4.964
F20	Avg	−3.2863	−3.32	−2.8029
	SD	0.057431	4.68×10^{-16}	0.38047
	Med	−3.322	−3.32	−3.0014
	Worst	−3.2031	−3.32	−1.9211
F21	Avg	−10.1532	−10.1	−3.0066
	SD	0	0.078	2.2011
	Med	−10.1532	−10.2	−3.1737
	Worst	−10.1532	−9.91	−0.4982
F22	Avg	−10.4029	−10.4	−4.1502
	SD	2.29×10^{-16}	2.43×10^{-16}	2.2208
	Med	−10.4029	−10.4	−4.707
	Worst	−10.4029	−10.4	−0.90936
F23	Avg	−10.5364	−8.45	−4.4381
	SD	1.78×10^{-15}	3.43	0.97997
	Med	−10.5364	−10.5	−4.6828
	Worst	−10.5364	−1.86	−2.8115

Avg: average; SD: standard deviation; Med: median; worst: worst of the best-so-far solution; NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

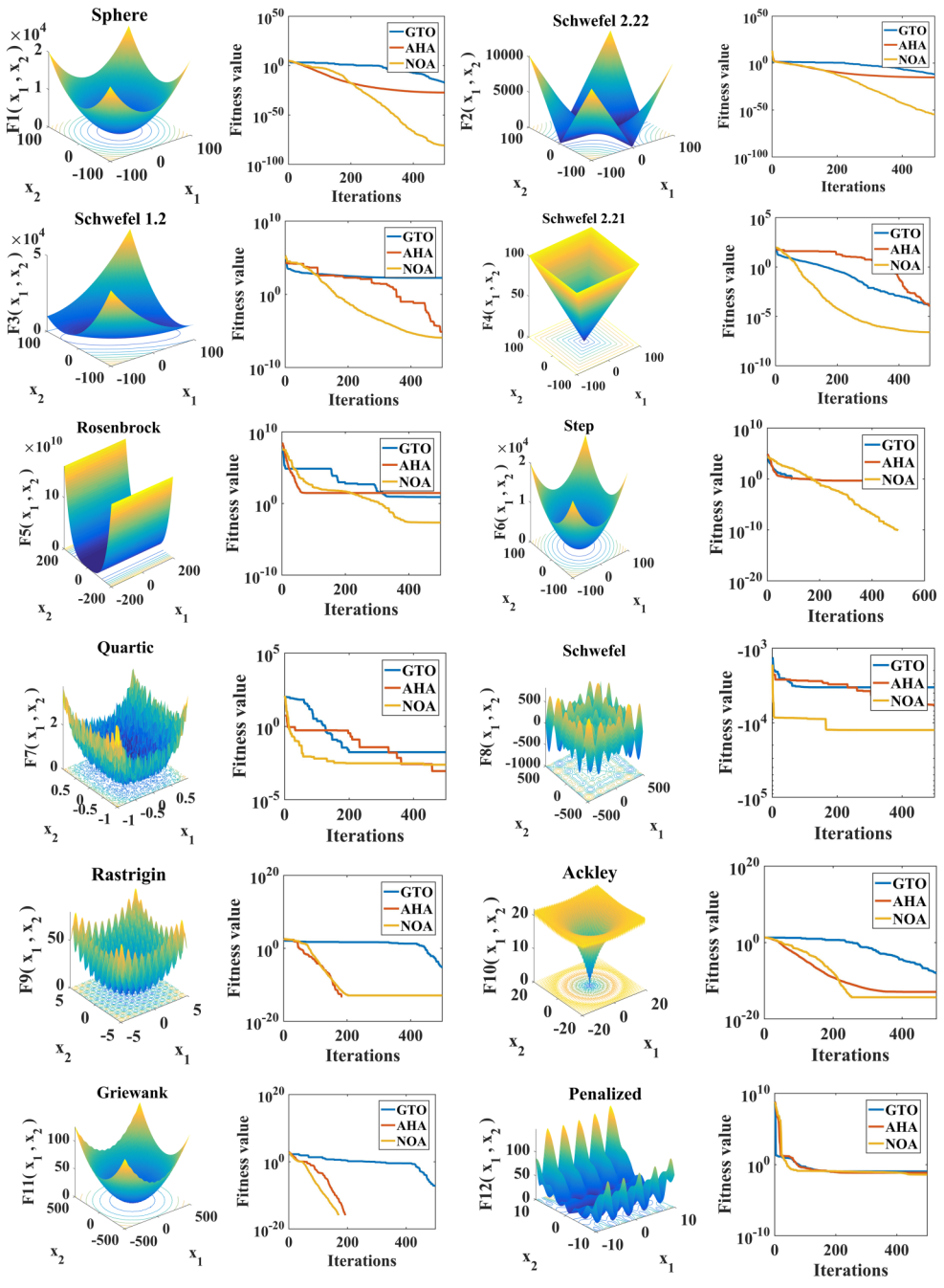


Figure 1. Benchmark function representation and convergence profile of NOA, GTO, and AHA. NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm. (continued)

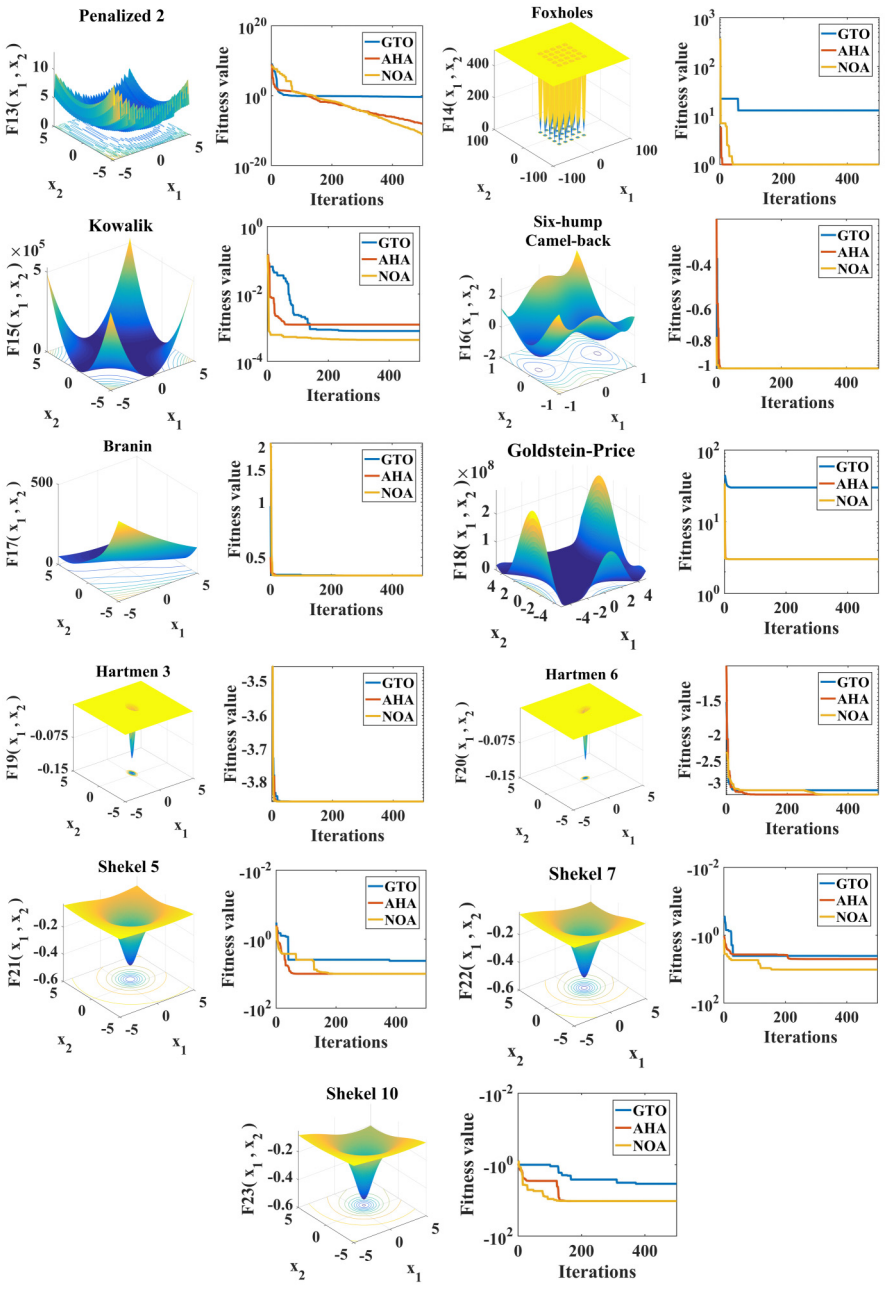


Figure 1. Continued.

NOA to solve the CM problem, the same problem has also been solved with other recent well-known optimization techniques such as gorilla troops optimization (GTO), and artificial hummingbird algorithm (AHA). These techniques are applied in the same environment, enabling a fair comparison among different optimization methods for solving the CM problem.

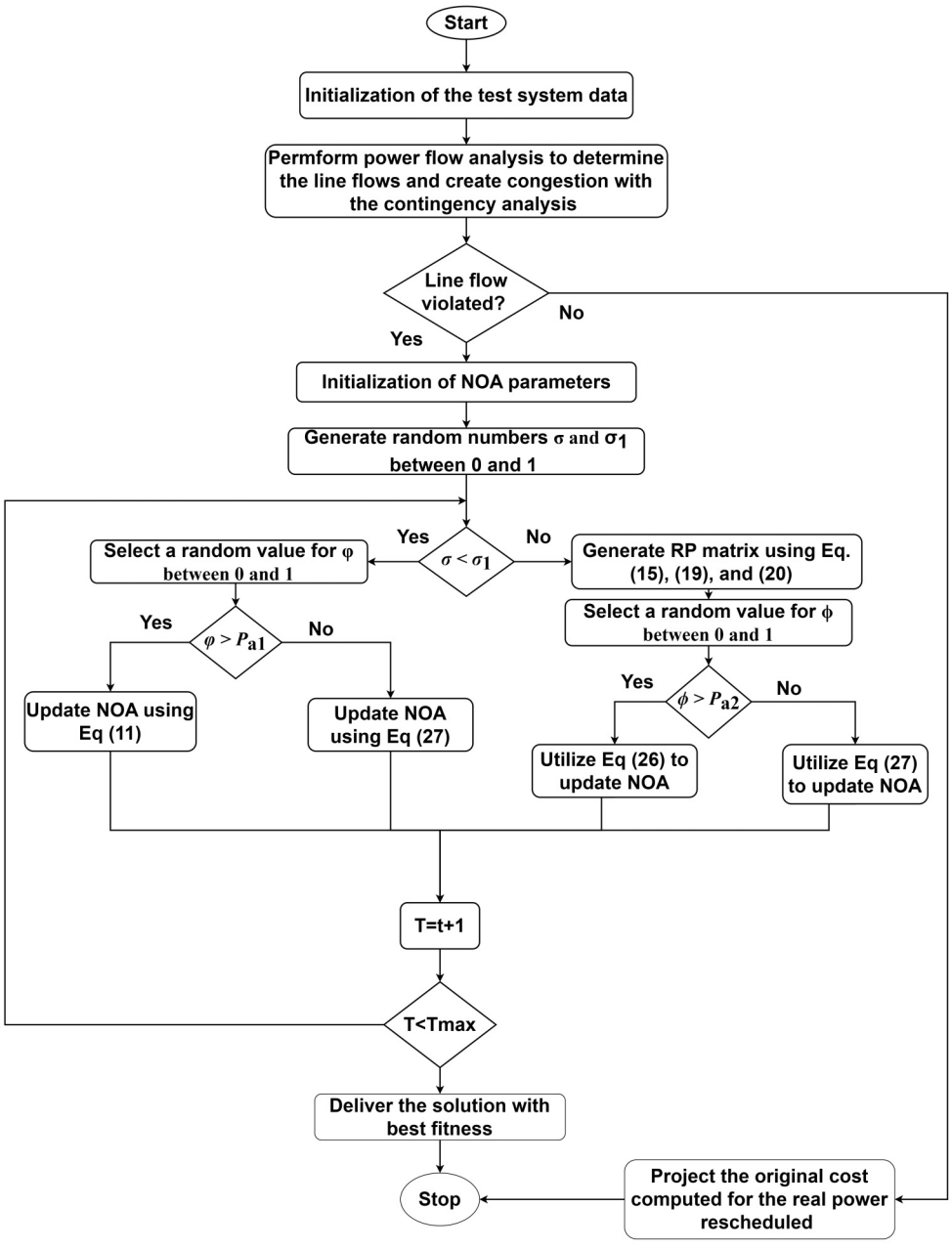


Figure 2. Flowchart for CM with NOA.
NOA: nutcracker optimizer algorithm; CM: congestion management

The software used in this study is MATLAB 2016a and implemented on a system with an Intel Core i7 processor running at a clock speed of 2.4 GHz and equipped with 8 GB of RAM. Table 3 provides specific information about the simulated cases conducted on the test systems. In this

Table 2. Computational complexity time (NOA) in seconds.

Dimension (D)	T_0	T_1	T_2	$(T_2-T_1)/T_0$
20	0.7539	7.8295	38.1602	34.3647
40	0.7539	44.3275	107.3527	75.1853
50	0.7539	106.2573	214.6285	132.5194

Table 3. Test cases considered for the congestion management (CM) problem.

Cases	Test system	Contingency scenario considered
1	IEEE 30-bus system (Sharma et al., 2018)	Tripping of lines 1 and 2
2	IEEE 118-bus system (Dash and Mohanty, 2015)	Tripping of lines 8–5 and load increment to 57% at buses 11–20

Table 4. Line power flow details for the congested line.

	Congested lines	Power flow (MW)	Power flow limit (MW)	Excess power flow (MW)	% overload
Outage of lines 1 and 2	1–3	147.46	130	17.46	13.43
	6–8	136.29	130	6.29	4.84

research, line overloads have been created considering two cases, firstly, by decreasing the power flow capacity of the capacity of the lines, and secondly, by simulating line outage(s) for contingency purposes.

To find the optimal parameter values, multiple trial runs have been conducted for solving the CM problem on all three test systems under study. The best results obtained from these trials are presented. Through these trial runs, the population size of 40 has been considered to be sufficient for the CM problem in this particular study.

IEEE 30-bus system

The modified IEEE 30-bus test system has been considered for the evaluation of the CM problem. It consists of 41 transmission channels with six generator buses and 24 load buses. The configuration of this system is depicted in Figure 2. The test case that has been considered for the IEEE 30-bus system is listed in Table 3. The details of the congested power flow scenario are given in Table 4.

In this case, the congestion has been created by an outage of line 1, between buses 1 and 2 (Figure 3). This leads to congestion in line 2 (connecting buses 1 and 3) and line 4 (connecting buses 6 and 8) and is represented in Table 4. To ensure secure functioning, counteractive measures are necessary to relieve these overloaded lines. The NOA has been employed to minimize the congestion cost, with the optimal generator output. The congestion costs obtained using the NOA are presented in Table 5. These values are then compared with results from literature, such as PSO (Kamaraj, 2011), RSM (Kamaraj, 2011), SA (Kamaraj, 2011), FPA (Verma and Mukherjee, 2016b), CBA (Verma and Mukherjee, 2016b), ALO (Verma and Mukherjee, 2016b), TLBO (Verma et al., 2018) as well as recent algorithms like GTO, AHA. The optimal total congestion cost obtained through the NOA-based approach is determined to be 476.84 $/h.

Table 5. Comparative outcomes for IEEE 30 bus system with NOA.

	RSM (Kamaraj, 2011)	PSO (Kamaraj, 2011)	SA (Kamaraj, 2011)	FPA (Verma and Mukherjee, 2016b)	CBA (Verma and Mukherjee, 2016b)	ALO (Verma and Mukherjee, 2016b)	TLBO (Verma et al., 2018)	GTO [solved]	AHA [solved]	NOA [proposed]
Congestion cost ($/h)	716.25	538.95	719.86	519.62	482.02	480.04	494.66	526.79	503.15	476.84
Best congestion cost ($/h)	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	526.79	503.15	476.84
Worst congestion cost ($/h)	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	796.53	686.48	510.84
Line (1–3) power flow (post CM)	129.78	129.97	129.51	129.60	129.97	129.5	130.00	129.40	129.57	129.40
Line (6–8) power flow (post CM)	120.60	120.62	120.35	120.58	120.79	120.79	120.78	120.38	120.65	120.03
ΔP_1 (MW)	−8.808	−8.61	−9.076	−9.127	−8.694	−9.0880	−8.5876	−9.180	−8.975	−7.7642
ΔP_2 (MW)	2.647	10.40	3.133	14.14	13.917	15.0668	12.9855	12.19	2.634	10.4476
ΔP_3 (MW)	2.953	3.03	3.234	0.206	0.016	0.0198	0.4598	0.206	3.08	0.3487
ΔP_4 (MW)	3.063	0.02	2.968	0.0188	0.109	0.0001	0.7289	0.018	1.869	0.6178
ΔP_5 (MW)	2.913	0.85	2.954	0.189	0.349	0.0002	0.0093	0.908	2.471	0.0074
ΔP_6 (MW)	2.952	0.01	2.443	1.013	0.317	0.0001	0.3988	1.117	1.082	0.2877
Total amount (MW)	23.33	22.93	23.80	24.703	23.402	24.1552	23.169	23.619	20.111	19.4734

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm; PSO: particle swarm optimization; FPA: flower pollination algorithm; RSM: random search method; SA: simulated Annealing; CBA: chaotic bat algorithm; ALO: ant lion optimizer; TLBO: teaching learning based optimization.
[a]NR: not reported.

Representation of the comparative congestion costs is depicted in Figure 4. Figure 5 illustrates the real power rescheduling for different generators achieved by the comparative algorithms. The convergence profiles of GTO, AHA, and NOA, are presented in Figure 6. The iterative values for 30 runs with 500 iterations and the box plot are given in Figures 7 and 8.

Additionally, the total system loss is reduced to 12.047 MW after CM, compared to the initial value of 16.023 MW during congestion. The bus voltages after CM using the NOA approach are shown in Figure 9, where it is evident that all bus voltages remain within acceptable limits.

Statistical test for NOA on IEEE 30-bus system. To assess the effectiveness and resilience of the proposed NOA, an extensive examination has been conducted on the obtained results. This involved

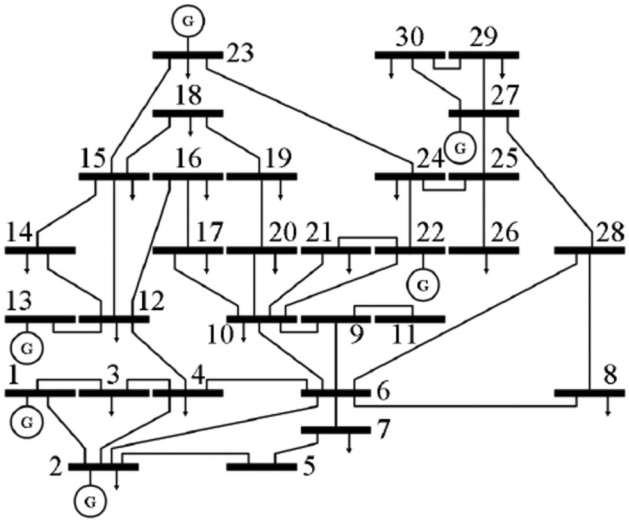


Figure 3. Single line representation of IEEE 30-bus system (Verma et al., 2018).

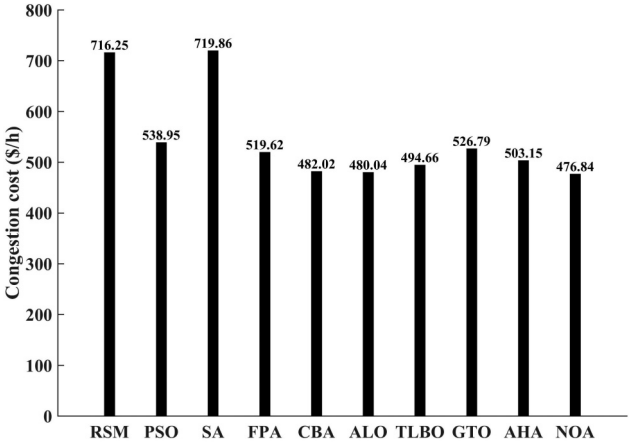


Figure 4. Comparative congestion cost with nutcracker optimizer algorithm (NOA) and other optimization algorithms.

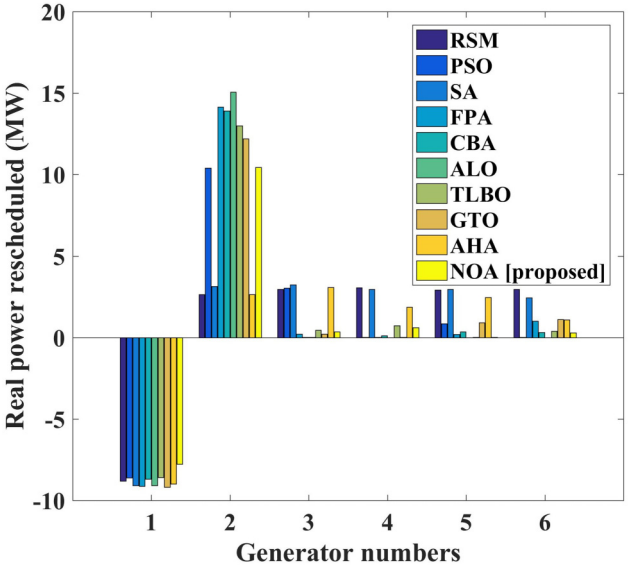


Figure 5. Comparative representation of the real power rescheduled with the optimization algorithms.

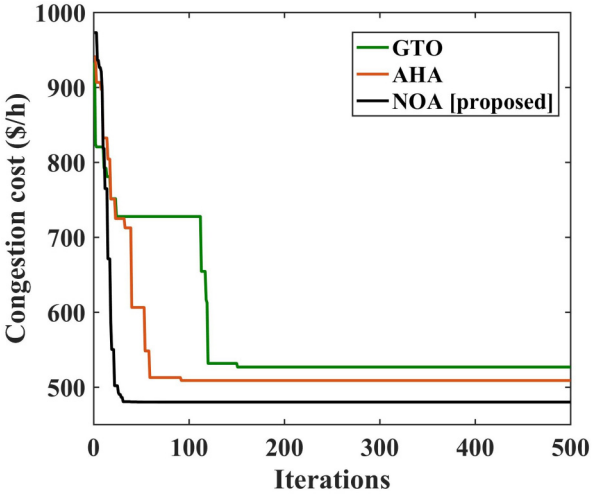


Figure 6. Convergence characteristics of nutcracker optimizer algorithm (NOA) and other optimization algorithms.

running all three optimization algorithms 30 times across various population sizes (number of search agents). According to Derrac et al. (2011), the authors recommended employing several non-parametric statistical tests within the realm of computational intelligence. Table 6 shows the cases with the best fitness function values. Notably, the NOA yielded the best result when utilizing 40 search agents. Table 7 presents the minimum, maximum, mean, and standard deviations of fitness function values obtained with GTO, AHA, and NOA when using 40 search agents.

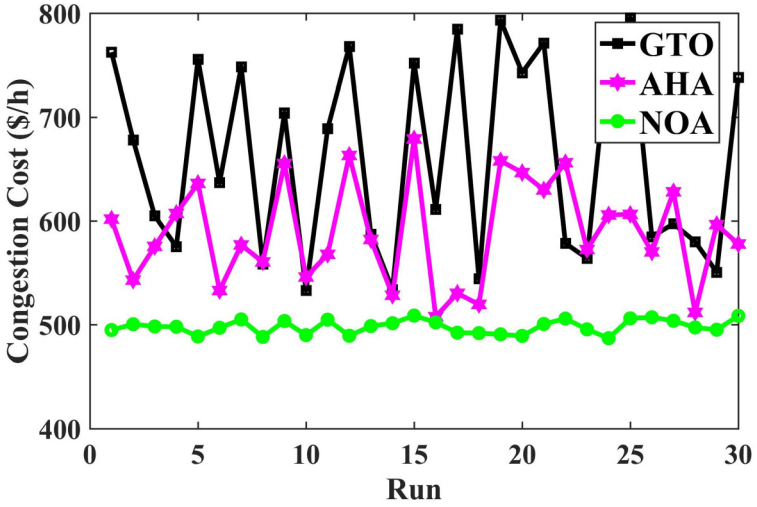


Figure 7. Fitness function values for 30 runs with 500 iterations (IEEE 30 bus).

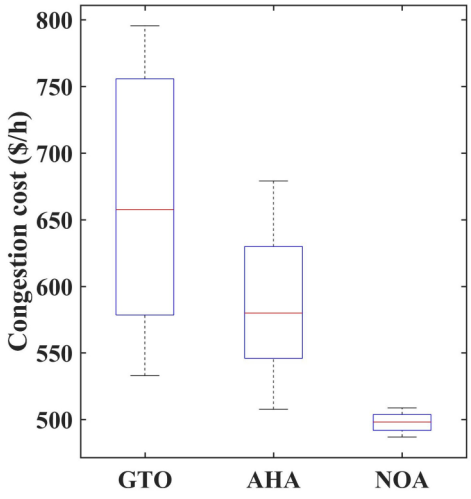


Figure 8. Box plot for 30 trail runs (IEEE 30 bus).

Furthermore, statistical analyses including Friedman, and Quade tests have been conducted to validate the optimization algorithms. Let H_0 denote the hypothesis stating no difference among the three methods, while H_1 represents the alternative hypothesis suggesting differences (Sheskin, 2003). A significance level of $\alpha = 0.05$ has been chosen. Table 8 presents the recorded *p*-values from Wilcoxon's paired sign rank test. The symbols +/−/∼ indicate whether the proposed NOA outperforms, underperforms, or is approximately equal to the compared algorithm, respectively. A *p*-value < 0.05 for all test cases refutes the null hypothesis, affirming the validity of Wilcoxon's signed-rank test for the proposed algorithm (Sheskin, 2003).

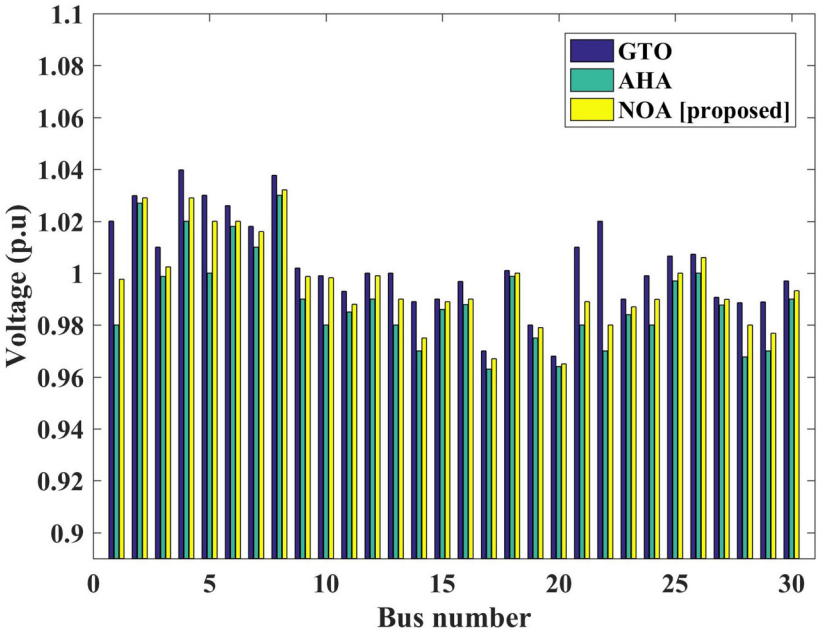


Figure 9. Comparative bus voltage with nutcracker optimizer algorithm (NOA) and other optimization techniques.

Table 6. The fitness function value is delivered by the optimization algorithm for different search agents.

	Search agents			
Algorithms	20	30	40	50
GTO	531.68	528.90	526.79	524.25
AHA	509.16	507.99	503.15	502.17
NOA	483.36	479.56	476.84	474.75

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

Table 7. Fitness function values delivered by GTO, AHA, and NOA for 40 search agents

Algorithms	Min	Max	Avg	SD
GTO	526.79	796.53	662.905	12.6942
AHA	503.15	686.48	589.049	9.0678
NOA	476.84	510.84	497.966	4.7358

Avg: average; SD: standard deviation; Med: median; worst: worst of the best-so-far solution; NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

Analysis of variance relies on the principle of total variance. This approach involves breaking down the observed variance in a specific variable into segments attributed to various sources of variation. The critical F-value is determined with the chosen significance level ($\alpha = 0.05$). If the

Table 8. *p*-values for the optimization algorithm.

Algorithm	*p*-value	+/−/∼
GTO versus NOA	1.7344×10^{-6}	+
AHA versus NOA	1.9209×10^{-6}	+

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

Table 9. Analysis of variance for the optimization algorithms.

Source of variation	Sum of square	Degrees of freedom	Mean square	*F*-ratio	5% *F*-limit
Between techniques	3.5257×10^{7}	5–1 = 4	9.0420×10^{6}	14.6751	$F\ (4, 25) = 2.8696$
Within techniques	1.3718×10^{7}	30–5 = 25	5.7279×10^{5}		

Table 10. Representation of average error for GTO, AHA, and NOA.

Search agents	GTO	AHA	NOA
20	23.2889	32.43046	2.74084
30	45.87211	13.86326	1.69437
40	38.75376	2.654998	1.49064
50	28.10499	26.79083	2.76185

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

Table 11. Friedman test analysis data.

	Friedman ranks		
Search agents	GTO	AHA	NOA
20	4	2	1
30	4	2	1
40	4	2	1
50	4	2	1
Avg. Rank	4	2	1
F-statistic	16		
P-value	0.0020		

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

calculated F-ratio exceeds the critical *F*-limit, the null hypothesis is rejected. Additionally, Table 9 indicates that the computed *F*-value falls below the tabulated *F*-value at a 5% significance level, where the degrees of freedom are specified as 4 and 25.

To conduct the Friedman and Quade tests, initially, an average error table is highlighted in Table 10. Subsequently, Friedman ranks and the *F*-statistic (also known as the χ^2 value) are computed and presented in Table 11. The method with the lowest rank is deemed superior, clearly indicating the efficacy of NOA. The computed χ^2 value of 16 from Table 11 exceeds

Table 12. Representation of Quade test analysis for GTO, AHA, and NOA.

	Quade ranks		
Search agents	GTO	AHA	NOA
20	4	−8	−6
30	4	−2	−3
40	4	−6	−2
50	4	2	−1
Sum Rank	4	2	−12
Q-statistic	15		
F-value	$1.3970 \times 10^{-0.3}$		

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

Table 13. Details of congested lines (118 bus).

	Congested lines (MW)	Power flow (MW)	Line limit (MW)	Excess power flow (MW)	% overload
Tripping of lines 8–5 with an increase in load at buses 11–20 by 57%	16–17	209.14	175	34.14	12.83
	30–17	568.73	500	13.16	37.6
	8–30	380.59	175	205.59	54.01

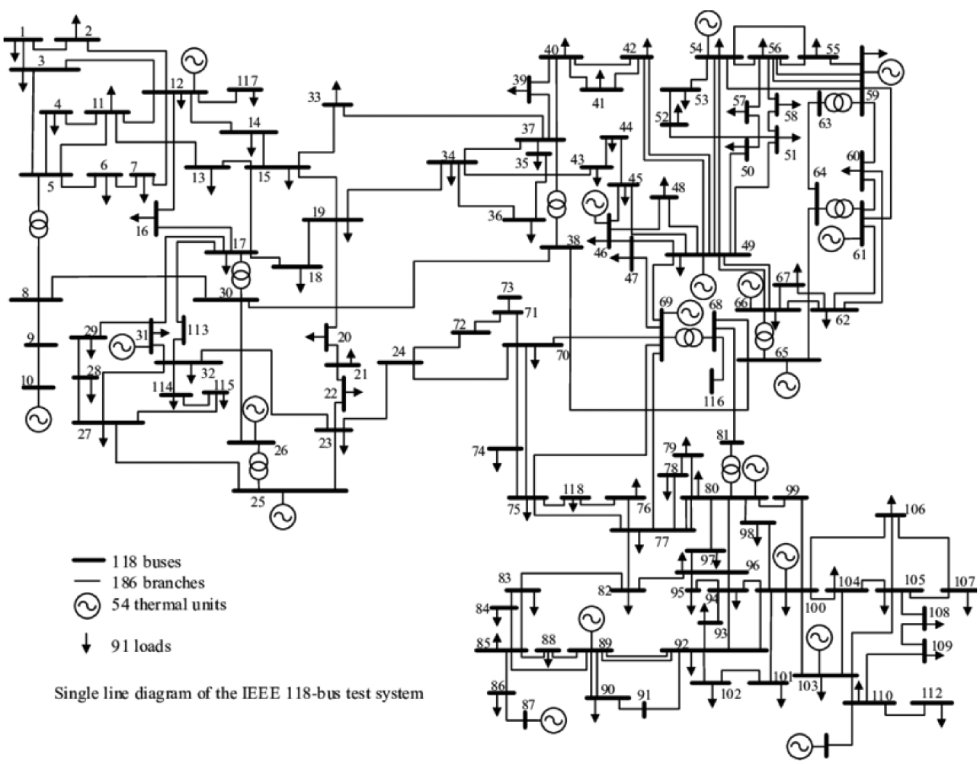


Figure 10. Single line diagram of 118-bus system (Wang et al., 2022).

Table 14. Comparative outcomes for IEEE 118-bus system with NOA.

	RCGA (Kamaraj, 2011)	EP (Kamaraj, 2011)	DE (Kamaraj, 2011)	HPSO (Verma and Mukherjee, 2016b)	PSO (Verma and Mukherjee, 2016b)	CBA (Verma and Mukherjee, 2016b)	ALO (Verma et al., 2018)	GTO [solved]	AHA [solved]	NOA [proposed]
Congestion cost ($/h)	17,693	42,886	16,080	17,365	17,742	15,724	15, 448	1528.07	1438.90	1389.56
Best congestion cost ($/h)	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	1528.07	1438.90	1389.56
Worst congestion cost ($/h)	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	NR[a]	2100.48	1845.78	1420.56
Lines (16 and 17) power flow (post CM)	174.26	173.89	174.37	174.62	174.15	174.09	174.20	174.23	174.87	173.18
Lines (30–17) power flow (post CM)	499.09	498.14	496.88	495.42	496.09	499.36	497.94	496.36	497.30	495.42
Line (8–30) power flow (post CM)	174.68	174.54	174.20	174.59	174.82	174.31	174.76	174.48	174.00	174.02
ΔP_{G1}	26.274	38.11	6.4705	11.59	24.502	0.0124	0.0177	13.317	4.8963	11.934
ΔP_{G2}	0	0	0	0	0	0.0455	0.0362	0.0087	0.0051	9.6558
ΔP_{G3}	31.422	63.06	12.295	6.568	1.0259	0.1476	17.266	0.7567	0.0035	1.665
ΔP_{G4}	11.349	−7.48	−0.056	−0.046	1.3148	−0.0261	−0.0031	83.455	−0.0005	0
ΔP_{G5}	−219.3	−200.6	−208.1	−207.7	−21.82	−207.8	−207.83	−0.2312	0.0237	0
ΔP_{G6}	101.42	48.58	105.32	80.126	105.32	105.31	105.32	1.4716	0.0001	113.40
ΔP_{G7}	0	0	0	0	0	0.0086	0.002	5.777	0.0452	101.12
ΔP_{G8}	0.0029	49.17	1.9946	6.0059	16.815	−0.0081	4.6839	22.78	0.0023	0
ΔP_{G9}	8.0582	64.056	1.2884	−0.239	23.665	0.0095	2.597	−0.125	0.2974	4.409
ΔP_{G10}	0	0	0	0	0	0.003	0.0042	3.0727	0.0223	15.742

(continued)

Table 14. Continued.

	RCGA (Kamaraj, 2011)	EP (Kamaraj, 2011)	DE (Kamaraj, 2011)	HPSO (Verma and Mukherjee, 2016b)	PSO (Verma and Mukherjee, 2016b)	CBA (Verma and Mukherjee, 2016b)	ALO (Verma et al., 2018)	GTO [solved]	AHA [solved]	NOA [proposed]
ΔP_{G11}	114.69	121.51	49.697	121.51	121.51	0.1359	0.0915	1.5325	0.0019	0
ΔP_{G12}	116.31	25.918	113.29	108.06	116.31	0.1555	28.186	2.5903	0.0276	0
ΔP_{G13}	7.4155	85.370	0.8125	10.752	3.0092	0.0074	−0.0024	3.7627	0.1654	−10.03
ΔP_{G14}	−0.0019	32.857	2.1836	−0.536	0.0177	0.0041	0.0678	12.6412	0.2261	11.165
ΔP_{G15}	28.095	31.616	89.669	11.254	8.8665	89.701	89.765	0.4368	46.1738	−0.5718
ΔP_{G16}	1.7508	39.071	0.7819	0.0731	7.1475	0.0006	0.0025	−0.4615	0.0002	−207.3
ΔP_{G17}	0.0004	37.552	−0.041	−0.394	0.4916	0.0884	0.004	3.0092	0.0006	104.93
ΔP_{G18}	0	0	0	0	0	0.0082	0.0426	0.0177	2.3777	0.8256
ΔP_{G19}	0	0	0	0	0	0.0022	0.0036	8.8665	116.12	0.0023
ΔP_{G20}	1.7478	89.001	−0.089	4.3096	6.0792	0	0.1504	7.1475	0.0001	1.6977
ΔP_{G21}	18.528	96.075	102.85	77.661	19.359	22.358	25.114	0.4916	0.2949	2.1221
ΔP_{G22}	33.426	93.828	7.3786	72.901	36.798	0.4626	8.6749	28.186	69.115	8.3874
ΔP_{G23}	0	0	0	0	0	0.0236	−0.0141	−0.0024	0.0140	13.317
ΔP_{G24}	0	0	0	0	0	0.0003	0.0059	0.0678	−0.0273	0.0087
ΔP_{G25}	0	0	0	0	0	9.0881	17.745	89.765	−0.0004	0.7567
ΔP_{G26}	0	0	0	0	0	10.296	8.1664	0.0025	0.9068	−0.0018
ΔP_{G27}	0	0	0	0	0	0.0101	−0.0024	0.004	0.0748	28.095
ΔP_{G28}	0	0	0	0	0	3.3333	33.28	0.0426	42.198	1.7508
ΔP_{G29}	0	0	0	0	0	14.779	9.1758	0.0036	0.0004	0.0004
ΔP_{G30}	0	−450.9	4.904	0.6525	10.702	1.1904	0.6717	0.1504	0.0052	0
ΔP_{G31}	0	0	0	0	0	0.0015	0.0173	25.114	0.0001	0
ΔP_{G32}	0	0	0	0	0	−0.0016	0.0032	8.6749	16.815	1.7478
ΔP_{G33}	0	0	0	0	0	0.3726	0.0218	−0.0141	23.665	18.528
ΔP_{G34}	0	0	0	0	0	0.2975	0.0024	48.58	0	33.426
ΔP_{G35}	0	0	0	0	0	0.0656	0.1945	0	121.51	10.752
ΔP_{G36}	0	0	0	0	0	0.0048	0.0013	49.17	116.31	−0.536
ΔP_{G37}	0	0	0	0	0	137.81	39.667	64.056	3.0092	11.254
ΔP_{G38}	0	0	0	0	0	0.0491	0.0372	0	0.0177	0.0731

(continued)

Table 14. Continued.

	RCGA (Kamaraj, 2011)	EP (Kamaraj, 2011)	DE (Kamaraj, 2011)	HPSO (Verma and Mukherjee, 2016b)	PSO (Verma and Mukherjee, 2016b)	CBA (Verma and Mukherjee, 2016b)	ALO (Verma et al., 2018)	GTO [solved]	AHA [solved]	NOA [proposed]
ΔP_{G39}	0	0	0	0	0	0.0145	2.7964	121.51	8.8665	−0.394
ΔP_{G40}	0	0	0	0	0	−0.003	0.0031	25.918	7.1475	0
ΔP_{G41}	0	0	0	0	0	0.0317	0.0033	85.370	0.4916	0
ΔP_{G42}	0	0	0	0	0	0.0835	0.0151	32.857	0.004	4.3096
ΔP_{G43}	0	0	0	0	0	0.0413	0.4569	1.2884	0.0426	77.661
ΔP_{G44}	0	0	0	0	0	0.0055	0.0073	0	0.0036	72.901
ΔP_{G45}	0	0	0	0	0	0.0035	0.196	49.697	0.1504	4.9421
ΔP_{G46}	719.8998	1574.9	707.3281	720.4816	717.7839	24.224	3.2218	113.29	205.114	71.938
ΔP_{G47}	0	0	0	0	0	0.0019	−0.0015	0.8125	8.6749	0.9656
ΔP_{G48}	0	0	0	0	0	−0.012	−0.0013	2.1836	−0.0141	−0.092
ΔP_{G49}	0	0	0	0	0	0.0096	0.0039	89.669	0.0059	1.53
ΔP_{G50}	0	0	0	0	0	0.0056	0.0012	0.7819	17.745	0.0023
ΔP_{G51}	0	0	0	0	0	102.026	100.433	116.31	8.1664	0.3837
ΔP_{G52}	0	0	0	0	0	0.0277	−0.0045	7.4155	−0.0024	22.78
ΔP_{G53}	0	0	0	0	0	0.0032	7.3507	−0.0019	33.28	−0.125
ΔP_{G54}	0	0	0	0	0	−0.0001	0.0009	28.095	9.1758	3.0727
Total amount (MW)	1439.6913	2149.654	1414.5492	1440.8597	1242.5373	730.1113	713.3677	1160.9855	863.2383	986.302

RCGA: Real coded GA; DE: differential evolution; HPSO: hybrid PSO; CBA: chaotic bat algorithm; ALO: ant lion optimizer.
NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm; PSO: particle swarm optimization.
[a]NR: not reported.

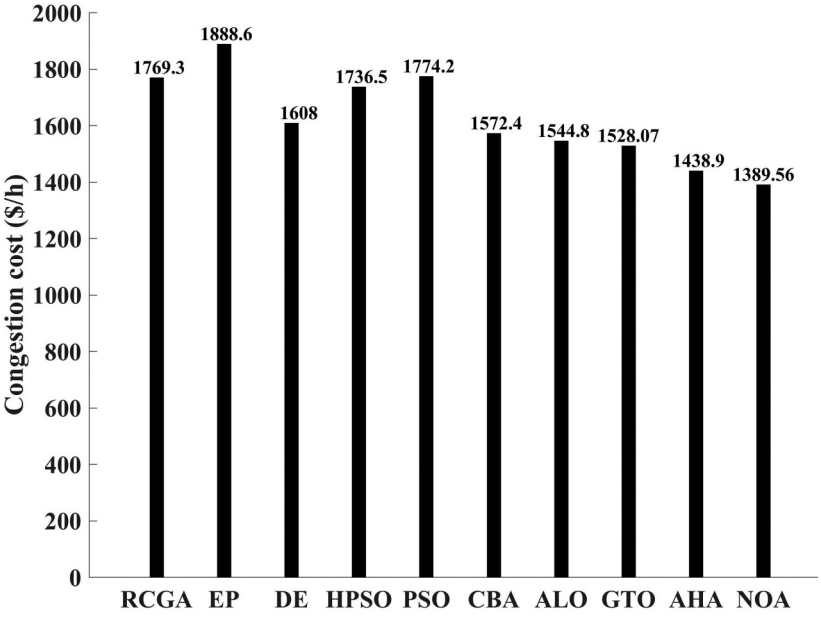


Figure 11. Comparative congestion cost with nutcracker optimizer algorithm (NOA) and other optimization algorithms.

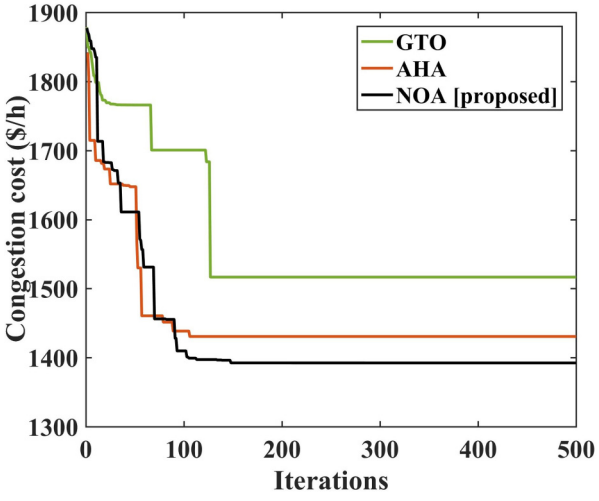


Figure 12. Convergence characteristics of nutcracker optimizer algorithm (NOA) and other optimization algorithms.

the critical value, and the resulting p-value of 0.0030, being < 0.05, leads to the rejection of the null hypothesis. This confirms the presence of significant disparities among the methods employed.

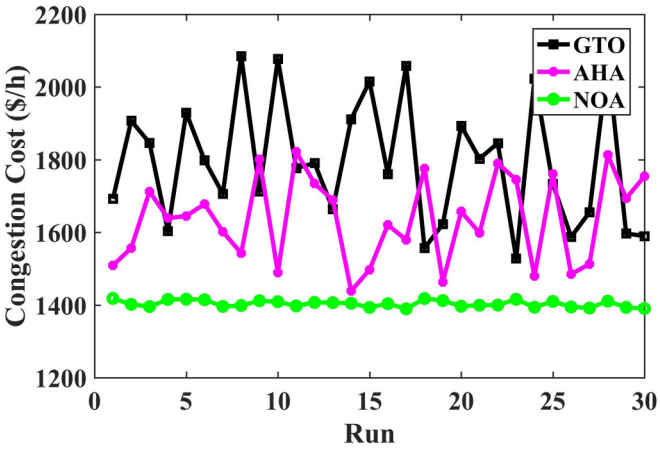


Figure 13. Fitness function values for 30 runs with 500 iterations (IEEE 118 bus).

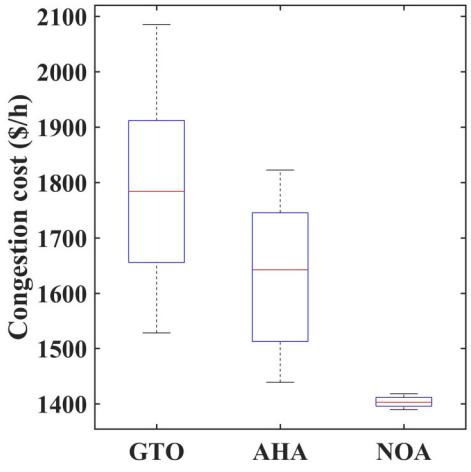


Figure 14. Figure 8 box plot for 30 trail runs (IEEE 118 bus).

Table 15. Fitness function value delivered by the optimization algorithm for different search agents (118 bus).

	Search agents			
Algorithms	20	30	40	50
GTO	1536.47	1534.09	1528.07	1524.27
AHA	1446.61	1442.17	1438.90	1436.79
NOA	1396.36	1391.56	1389.56	1380.58

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

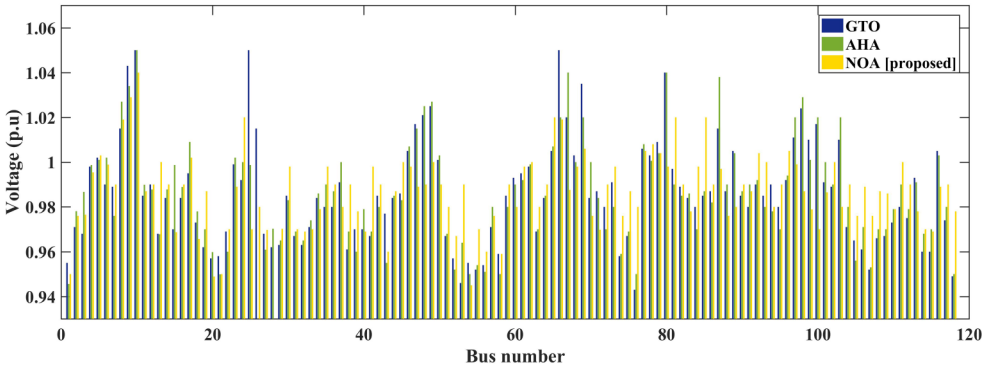


Figure 15. Comparative bus voltage with nutcracker optimizer algorithm (NOA) and other optimization techniques.

Table 16. Fitness function values delivered by GTO, AHA, and NOA for 40 search agents (118 bus).

Algorithms	Min	Max	Avg	SD
GTO	1528.07	2100.48	1602.05	14.9342
AHA	1438.90	1845.78	1549.49	12.3857
NOA	1389.56	1420.56	1436.66	3.8469

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

Table 17. *P*-values for the optimization algorithm (118 bus).

Algorithm	*p*-value	+/−/~
GTO versus NOA	1.6723×10^{-6}	+
AHA versus NOA	1.7318×10^{-6}	+

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

Table 18. Analysis of variance for the optimization algorithms (118 bus)

Source of variation	Sum of square	Degrees of freedom	Mean square	*F*-ratio	5% *F*-limit
Between techniques	3.6319×10^{7}	5–1 = 4	8.1479×10^{6}	15.5348	F (4, 25) = 3.4186
Within techniques	1.2187×10^{7}	30–5 = 25	4.5547×10^{5}		

Similar trends are evident in the Quade test results presented in Table 12, where NOA achieved the lowest rank. The calculated Q-statistic value of 15 surpasses the critical value referenced by Sheskin (2003), and the resulting p-value of 1.9209×10^{-6}, being < 0.05, supports the rejection of the null hypothesis H_0. This reinforces the effectiveness of NOA.

Table 19. Representation of average error for GTO, AHA, and NOA (118 bus).

Search agents	GTO	AHA	NOA
20	34.39910	43.54157	3.85195
30	56.98322	24.97437	2.70548
40	49.86487	3.765009	2.50175
50	39.215010	37.80194	3.87296

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

Table 20. Representation of FRIEDMAN test analysis for GTO, AHA, and NOA (118 bus).

	Friedman ranks		
Search agents	GTO	AHA	NOA
20	3	2	1
30	3	2	1
40	3	2	1
50	3	2	1
Avg. Rank	3	2	1
F-statistic	15		
P-value	0.0040		

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

Table 21. Representation of Quade test analysis for GTO, AHA, and NOA (118 bus).

	Quade ranks		
Search agents	GTO	AHA	NOA
20	4	−7	−7
30	4	−3	−2
40	4	−5	−3
50	4	3	−2
Sum Rank	16	−12	−14
Q-statistic	14		
F-value	$1.2851 \times 10^{-0.3}$		

NOA: nutcracker optimizer algorithm; GTO: gorilla troops optimization; AHA: artificial hummingbird algorithm.

IEEE 118-bus system

This test framework consists of 54 buses as generator buses, 186 transmission lines, and 64 load buses. The network topology for this test system can be found in Figure 10.

In this test system, contingency is established with the tripping of lines 8–5 and increasing the load demand by 1.57 times. Congested lines detail have been represented in Table 13. The NOA has been employed to settle the CM problem and obtain the solution. The detail of the solutions achieved with the NOA has been highlighted in Table 14. It can be seen that the line flows have been reduced below the maximum limits with the NOA and other competitive

optimization algorithms. The congestion cost and the real power rescheduled by the NOA, and other recent algorithms are summarized in Table 14. It is observed that the application of NOA and generator rescheduling have relieved the overloaded lines achieving a congestion cost of 1389.56 $/h, which is comparatively lower than the cost achieved with other techniques.

The congestion cost is represented in Figure 11. Convergence of the fitness function based on the NOA for this case is depicted in Figure 12. System loss before CM was 247.968 MW, whereas post CM it has been reduced to 168.539 MW. Furthermore, tests have shown that NOA consistently provides the best optimal solution even for large systems. The iterative values for 30 runs with 500 iterations and the box plot are given in Figures 13 and 14. The bus voltages obtained post-CM are represented in Figure 15, demonstrating that the voltages are within normal specified limits.

Statistical test for NOA on IEEE 118-bus system. The fitness obtained with GTO, AHA, and NOA is highlighted in Table 15. Notably, the NOA yielded the best result when utilizing 40 search agents. Table 16 presents the minimum, maximum, mean, and standard deviations of fitness function values. Various statistical analyses, including Wilcoxon's signed-rank test, Friedman, and Quade tests, were undertaken to assess the algorithms. H_0 represents the hypothesis asserting no distinctions among the three methods, while H_1 suggests otherwise (Verma and Mukherjee, 2016a). A significance level of $\alpha = 0.05$ was selected for the analyses. The *p*-values obtained from Wilcoxon's paired sign rank test are presented in Table 17. The symbols +/−/∼ indicate whether the proposed NOA performs better, worse, or comparably to the compared algorithm, respectively. A *p*-value < 0.05 across all test cases negates the null hypothesis, confirming the credibility of Wilcoxon's signed-rank test for the proposed algorithm (Verma and Mukherjee, 2016a).

Additionally, as depicted in Table 18, the computed *F*-value for both systems falls below the tabulated *F*-value at a 5% significance level, with degrees of freedom specified as 4 and 25. In conducting the Friedman and Quade tests, an average error table is established in resembling Table 19. Subsequently, Friedman ranks, and the *F*-statistic (also referred to as the χ^2 value) are computed and presented in Table 20. The method achieving the lowest rank is deemed superior, thereby highlighting the effectiveness of NOA. The calculated χ^2 value of 16 from Table 20 surpasses the critical value, and the resulting *p*-value of 0.0030, falling below 0.05, leads to the rejection of the null hypothesis, indicating significant disparities among the employed methods. Similar patterns are observable in the Quade test outcomes showcased in Table 21, where NOA secures the lowest rank. The calculated *Q*-statistic value of 15 exceeds the referenced critical value by Verma and Mukherjee (2016a), and the resulting *p*-value of 1.9209×10^{-6}, being below 0.05, supports the rejection of null hypothesis H_0, thus reinforcing the efficacy of NOA.

Conclusions

This research introduces a novel approach for power system CM within the power market, utilizing a generation rescheduling method. The contribution can be stated as the formulation of a novel CM model to minimize congestion costs through optimal rescheduling of generators' real power, utilizing NOA for determining optimal power rescheduling patterns. This model's efficacy is validated on power system standard test systems, IEEE 30 bus, and IEEE 118 bus. Comparative performance analysis is conducted, assessing congestion cost reduction, bus voltage profile improvement, and system losses, demonstrating the superiority of the proposed CM model with NOA over contemporary optimization techniques. The method employs a unique NOA and accounts for contingencies like line outages and sudden load variations. It aims to alleviate congestion in overloaded lines

by adjusting the power generation schedules while minimizing the congestion cost. The NOA technique has been successfully applied to IEEE 30- and IEEE 118-bus systems for CM purposes. Comparisons are made between the outcomes generated by the proposed NOA and various established algorithms documented in recent literature. Results indicate that the power rescheduled with NOA incurs lower congestion costs as compared to other similar techniques. Furthermore, applying CM reduces the overall system loss.

Consequently, the NOA offers a fresh and efficient approach to resolving power system CM challenges within a deregulated framework. Considering future approaches, the suggested approach for power system congestion management, employing NOA, has the potential for expansion towards optimizing power system operations, encompassing power dispatch and load scheduling strategies. It can also be utilized to examine the optimal positioning and control of power delivery from renewable energy systems and distributed generation, offering a comprehensive analysis of the proposed methodology.

Further, researchers dealing with large-scale power system issues might find the NOA a valuable optimization tool based on its demonstrated effectiveness. Research on power system transmission line congestion management using optimization techniques encounters several challenges. These include the complexity of power networks, data availability and accuracy issues, long computational times, the need for adaptable algorithms, and the integration of renewable energy sources. Overcoming these limitations is essential for enhancing the effectiveness and practicality of optimization approaches in real-world applications. Further, the researchers dealing with large-scale power system issues might find the NOA a valuable optimization tool based on its demonstrated effectiveness.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Kaushik Paul https://orcid.org/0000-0002-4022-0038

References

Abdel-Basset M, Mohamed R, Jameel M, et al. (2023) Nutcracker optimizer: A novel nature-inspired metaheuristic algorithm for global optimization and engineering design problems. *Knowledge-Based Systems* 262: 110248.

Agrawal A, Pandey SN, Srivastava L, et al. (2022) Hybrid deep neural network-based generation rescheduling for congestion mitigation in spot power market. *IEEE Access* 10: 29267–29276.

Alavijeh NM, Steen D, Le AT, et al. (2024) Capacity limitation based local flexibility market for congestion management in distribution networks: Design and challenges. *International Journal of Electrical Power & Energy Systems* 156: 109742.

Ansari MM, Guo C, Shaikh MS, et al. (2020) Planning for distribution system with grey wolf optimization method. *Journal of Electrical Engineering & Technology* 15: 1485–1499.

Bouhouras AS, Kelepouris NS, Koltsaklis N, et al. (2024) Congestion management in coupled TSO and DSO networks. *Electric Power Systems Research* 229: 110145.

Chakravarthi K, Bhui P, Sharma NK, et al. (2022) Real time congestion management using generation re-dispatch: modeling and controller design. *IEEE Transactions on Power Systems* 28: 2189–2203.

Dash S and Mohanty S (2015) Multi-objective economic emission load dispatch with nonlinear fuel cost and noninferior emission level functions for IEEE-118 bus system. *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*. IEEE, pp.1371–1376.

Dehnavi E, Afsharnia S, Akmal AAS, et al. (2022) A novel congestion management method through power system partitioning. *Electric Power Systems Research* 213: 108672.

Dehnavi E, Akmal AAS and Moeini-Aghtaie M (2024) A novel day-ahead and real-time model of transmission congestion management using uncertainties prioritizing. *Electrical Engineering* 24: 1–14.

Derrac J, García S, Molina D, et al. (2011) A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation* 1: 3–18.

Dolatabadi SH, Ghorbanian M, Siano P, et al. (2020) An enhanced IEEE 33 bus benchmark test system for distribution system studies. *IEEE Transactions on Power Systems* 36: 2565–2572.

Gupta AR, Bharatbhai NM and Bhadoriya JS (2023) Performance analysis of distribution system with PV penetration using flower pollination algorithm. *Journal of the Institution of Engineers (India): Series B* 104: 441–459.

Haq FU, Bhui P and Chakravarthi K (2022) Real time congestion management using plug in electric vehicles (PEV's): A game theoretic approach. *IEEE Access* 10: 42029–42043.

Hobbie H, Mehlem J, Wolff C, et al. (2022) Impact of model parametrization and formulation on the explorative power of electricity network congestion management models—Insights from a grid model comparison experiment. *Renewable and Sustainable Energy Reviews* 159: 112163.

Kamaraj N (2011) Transmission congestion management using particle swarm optimization. *J Electrical Systems* 7: 54–70.

Liang JJ, Qu BY and Suganthan PN (2013) Problem definitions and evaluation criteria for the CEC 2014 special session and competition on single objective real-parameter numerical optimization. *Computational Intelligence Laboratory, Zhengzhou University, Zhengzhou China and Technical Report, Nanyang Technological University, Singapore* 635: 2014.

Mahajan V, Prajapati VK and Mudagal S (2022) Review of congestion management in deregulated power system. *Deregulated Electricity Structures and Smart Grids* 32: 259–289.

Majeed Rashid Zaidan SIT (2022) Emergency congestion management of power systems by static synchronous series compensator. *Indonesian Journal of Electrical Engineering and Computer Science* 25: 1258–1265.

Mishra RN, Yadav A and Singh MP (2022) Artificial intelligence techniques based congestion management in restructured power systems: a review. *2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*. IEEE, pp.1–7.

Ogunwole EI and Krishnamurthy S (2022) Transmission congestion management using generator sensitivity factors for active and reactive power rescheduling using particle swarm optimization algorithm. *IEEE Access* 10: 122882–122900.

Paul K (2022a) Modified grey wolf optimization approach for power system transmission line congestion management based on the influence of solar photovoltaic system. *International Journal of Energy and Environmental Engineering*: 1–17.

Paul K (2022b) Multi-objective risk-based optimal power system operation with renewable energy resources and battery energy storage system: A novel hybrid modified grey wolf optimization–sine cosine algorithm approach. *Transactions of the Institute of Measurement and Control*: 01423312221079962.

Paul K, Kumar N, Agrawal S, et al. (2019) Optimal rescheduling of real power to mitigate congestion using gravitational search algorithm. *Turkish Journal of Electrical Engineering & Computer Sciences* 27: 2213–2225.

Paul K, Sinha P, Mobayen S, et al. (2022) A novel improved crow search algorithm to alleviate congestion in power system transmission lines. *Energy Reports* 8: 11456–11465.

Roustaei M, Letafat A, Sheikh M, et al. (2022) A cost-effective voltage security constrained congestion management approach for transmission system operation improvement. *Electric Power Systems Research* 203: 107674.

Saravanan C and Anbalagan P (2022) An intelligent hybrid technique for optimal generator rescheduling for congestion management in a deregulated power market. *Journal of Intelligent & Fuzzy Systems* 43: 1–15.

Sarwar M, Siddiqui AS, Ghoneim SS, et al. (2022) Effective transmission congestion management via optimal DG capacity using hybrid swarm optimization for contemporary power system operations. *IEEE Access* 10: 71091–71106.

Shaikh MS, Ansari MM, Jatoi MA, et al. (2020) Analysis of underground cable fault techniques using MATLAB simulation. *Sukkur IBA Journal of Computing and Mathematical Sciences* 4: 1–10.

Shaikh MS, Hua C, Hassan M, et al. (2022a) Optimal parameter estimation of overhead transmission line considering different bundle conductors with the uncertainty of load modeling. *Optimal Control Applications and Methods* 43: 652–666.

Shaikh MS, Hua C, Jatoi MA, et al. (2021a) Application of grey wolf optimisation algorithm in parameter calculation of overhead transmission line system. *IET Science, Measurement & Technology* 15: 218–231.

Shaikh MS, Hua C, Jatoi MA, et al. (2021b) Parameter estimation of AC transmission line considering different bundle conductors using flux linkage technique. *IEEE Canadian Journal of Electrical and Computer Engineering* 44: 313–320.

Shaikh MS, Hua C, Raj S, et al. (2022b) Optimal parameter estimation of 1-phase and 3-phase transmission line for various bundle conductor's using modified whale optimization algorithm. *International Journal of Electrical Power & Energy Systems* 138: 107893.

Shaikh MS, Raj S, Babu R, et al. (2023) A hybrid moth–flame algorithm with particle swarm optimization with application in power transmission and distribution. *Decision Analytics Journal* 6: 100182.

Shaikh MS, Raj S, Ikram M, et al. (2022c) Parameters estimation of AC transmission line by an improved moth flame optimization method. *Journal of Electrical Systems and Information Technology* 9: 25.

Sharma P, Batish N, Khan S, et al. (2018) Power flow analysis for IEEE 30 bus distribution system. *WSEAS Transactions on Power Systems* 13: 48–59.

Sharma V and Walde P (2022) A novel optimisation technique based on swarm intelligence for congestion management in transmission lines. *International Journal of Power and Energy Conversion* 13: 1–23.

Sheskin DJ (2003) *Handbook of Parametric and Nonparametric Statistical Procedures*. Western Connecticut, USA: Chapman and hall/CRC.

Srivastava J and Yadav NK (2021) Rescheduling-based congestion management by metaheuristic algorithm: Hybridizing lion and moth search models. *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields* 35: jnm2952.

Subramaniyan V and Gomathi V (2023) A soft computing-based analysis of congestion management in transmission systems. *Tehnički Vjesnik* 30: 274–281.

Thiruvel A, Thiruppathi S, Chidambararaj N, et al. (2023) Modern power system operations in effective transmission congestion management via optimal DG capacity using firefly algorithms. 2023 *9th International Conference on Electrical Energy Systems (ICEES)*. IEEE, pp.360–365.

Verma S and Mukherjee V (2016a) Firefly algorithm for congestion management in deregulated environment. *Engineering Science and Technology, an International Journal* 19: 1254–1265.

Verma S and Mukherjee V (2016b) Optimal real power rescheduling of generators for congestion management using a novel ant lion optimiser. *IET Generation, Transmission & Distribution* 10: 2548–2561.

Verma S, Saha S and Mukherjee V (2018) Optimal rescheduling of real power generation for congestion management using teaching-learning-based optimization algorithm. *Journal of Electrical Systems and Information Technology* 5: 889–907.

Wang X, Zhao T and Parisio A (2022) Frequency regulation and congestion management by virtual storage plants. *Sustainable Energy, Grids and Networks* 29: 100586.

Wangunyu IG, Murage DK and Kihato PK (2022) Power system congestion management by generator active power rescheduling using cuckoo search algorithm. *Proceedings of the Sustainable Research and Innovation Conference*, pp.159–164.

Zakaryaseraji M and Ghasemi-Marzbali A (2022) Evaluating congestion management of power system considering the demand response program and distributed generation. *International Transactions on Electrical Energy Systems* 2022: 1–13.

Zhou H, Lu L, Wei M, et al. (2023) Robust scheduling of a hybrid hydro/photovoltaic/pumped-storage system for multiple grids peak-shaving and congestion management. *IEEE Access* 12: 22230–22242.

Appendix

Table A1. Bid cost for the generators in the congestion management for IEEE 30-bus system (Sharma et al., 2018).

Generator number	1	2	3	4	5	6
Incremental price bids ($MW/h)	21	19	40	42	44	39
Decremental price bids ($MW/h)	19	18	37	36	34	38

Table A2. Bid cost for the generators in the congestion management for IEEE 118-bus system (Dash and Mohanty, 2015).

Generator number	1	4	6	8	10	12	15	18	19
Incremental price bids ($MW/h)	39	42	40	43	21	22	44	40	43
Decremental price bids ($/MWh)	37	34	37	38	16	17	34	37	35
Generator number	24	25	26	27	31	32	34	36	40
Incremental price bids ($MW/h)	44	26	22	40	44	23	42	46	42
Decremental price bids ($MW/h)	34	17	17	38	34	17	36	35	34
Generator number	42	46	49	54	55	56	59	61	62
Incremental price bids ($MW/h)	46	44	24	31	41	42	22	25	43
Decremental price bids ($MW/h)	37	34	17	29	35	36	16	15	35
Generator number	65	66	69	70	72	73	74	76	77
Incremental price bids ($MW/h)	64	65	68	69	71	72	73	75	76
Decremental price bids ($MW/h)	16	17	14	38	38	37	35	38	37
Generator number	80	85	87	89	90	91	92	99	100
Incremental price bids ($MW/h)	79	84	86	88	89	90	91	98	99
Decremental price bids ($MW/h)	14	37	16	35	37	37	35	36	35
Generator number	103	104	105	107	110	111	112	113	116
Incremental price bids ($MW/h)	34	17	36	35	22	44	40	22	44
Decremental price bids($MW/h)	16	37	38	37	36	17	37	32	34

Table A3. Notation list.

CC	congestion cost ($/h)	U_j	upper bounds
C_{inc}	incremental price bids ($/h)	L_j	lower bounds
C_{dec}	decremental price bids($/h)	ψ	randomly generated number [0, 1]
ΔP_g	adjusted power (MW)	$Z^t_{m,j}$	mean solution
P_{Gk}	active power (MW)	$\xi_{t+1(new)}$	random number [0, 1]
Q_{Gk}	reactive power (MVA)	$\overrightarrow{Z}_t$	new position
P_{Dk}	power demand (MW)	$\overrightarrow{Z}_{best}$	best position
θ_{kj}	admittance angle	τ	random number [0, 1]
δ_k	voltage angle	γ	Levey flight value
N_b	number of buses	RPs	reference points
N_g	number of generators	$N_{d_{t+1}}$	number of loads
α	convergence factor	$\overrightarrow{Z}$	current position

Cluster Computing
https://doi.org/10.1007/s10586-024-04599-w

A novel software defect prediction model using two-phase grey wolf optimisation for feature selection

Ruchika Malhotra[1] · Kishwar Khan[1]

Received: 8 March 2024 / Revised: 13 May 2024 / Accepted: 23 May 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
The process of accurately predicting software defects is highly crucial during the early period of software development before testing activities begin. A variety of computational methods have been constructed to achieve this based on static code metrics. However, one of the major issues in predictive modelling is the presence of redundant and irrelevant features in available datasets, which can lead to inaccuracies in the prediction model. Swarm optimization methods have shown excellent performance in Feature Selection (FS) issue mitigation and reduced the execution time of the prediction model. This study proposes a novel model for predicting software defects. This model utilizes a variant of Grey Wolf Optimiser as a wrapper-based feature selection method, paired with Synthetic Minority Oversampling Technique to balance the dataset, with the objective of maximizing the prediction efficiency of the learning model. The performance of the proposed model is assessed on 27 open-source datasets. The result findings show that the feature selection method improves prediction performance. Furthermore, the two-phase Grey Wolf Optimization-based feature selection with Random Forest classifier demonstrates superior efficacy on datasets compared to another benchmark model in handling the problem of FS. The results are also validated using statistical techniques.

Keywords Software defect prediction · Feature selection · Grey wolf optimisation · Synthetic minority oversampling technique (SMOTE) · Machine learning

1 Introduction

Software quality is a significant aspect of software engineering that needs to be considered when developing software products. The extent to which software conforms to customer needs and expectations is the software quality [1]. High-quality software products are more likely to meet user needs and expectations, resulting in greater user satisfaction. High-quality software is effective and efficient, with minimal defects. This can result in smoother business processes and better use of resources. Quality software development can reduce overall testing and maintenance efforts, resulting in lower costs. Reducing the number of faults through quality development practises might ultimately save money because correcting defects can be expensive. High-quality software can give organizations a competitive advantage in the marketplace. Customers are more likely to choose a software product that meets their needs and performs well. Standardizing software quality practices can improve overall software quality and promote consistency across organizations [2].

Additionally, by predicting software defects, software engineers can improve their overall development processes, as they can use the data generated by Software Defect Prediction (SDP) to identify parts of the code that are prone to defects and make changes to reduce the risk of future defects data [3]. SDP is the process of identifying those software components that are most likely to have defects based on current software metrics or features for prospective software releases [4]. Software characteristics may indicate software properties like complexity and the number of operators, and the connections between the features and the class may change. Some of these traits

✉ Kishwar Khan
kishwarkhan037@gmail.com

Ruchika Malhotra
ruchikamalhotra@dtu.ac.in

[1] Department of Software Engineering, Delhi Technological University, Delhi, India

Published online: 08 June 2024

might be more relevant to the class than others, while some of them might be redundant or not necessary. With Feature Selection (FS), the high-dimensional feature set can yield high correlation features. FS may select only those elements that are essential to the class. Including FS in SDP could thereby boost its performance [5]. The problems of multicollinearity and the "curse of dimensionality" are avoided by only using a few features. The total count of features present in the given dataset can be decreased with the use of FS and reduction techniques.

The overall performance of ML classifiers will suffer from high data dimensionality and the presence of redundant, noisy, and irrelevant data. These features may be selected to decrease dimensionality and improve ML algorithm performance. Additionally, FS helps researchers and programmers comprehend the data on its own and focuses on the most significant and relevant features. The performance of SDP systems can be improved by FS algorithms, according to several research publications in the literature.

Dealing with huge datasets can be challenging for the classification process due to their high dimensionality. Because of this, utilizing the most useful features possible while also optimizing classification accuracy, FS pre-processing is a must for lowering the dimensionality of datasets. The FS for classification issues based on wrapper techniques is addressed in this study using a novel Grey Wolf Optimizer (GWO) procedure, including two phases of mutation [6].

The most useful features from the actual dataset can be chosen using a variety of FS techniques. Wrapper, filter, and embedded FS techniques fall within this category. Wrapper methods evaluate the chosen subset of characteristics using a classification procedure.

The filter approach assesses the chosen features according to the different aspects of the dataset. The filter methods are more rapid than the wrapper approach because the wrapper approach serves no purpose in assessing the classifier accuracy. As opposed to the filter technique, the improved set of features which correspond with the classifier may be obtained through the wrapper technique. Low computational cost is a benefit of hybrid approaches; however, they are often machine learning-specific. The two objectives that FS must accomplish are reducing the size of the feature pool and greater accuracy in classification.

Finding an ideal subset of features is crucial in FS [7]. By analyzing the whole collection of features, the exhaustive search can create all potential subsets. Due to the fact that there are 2M subsets of features for a dataset with M features, this method is unfeasible for big datasets and has an extraordinarily high computational cost. In the last two decades, metaheuristics approaches have proven to be productive and effective in resolving complex and large-scale issues in applications based on Machine Learning (ML) and data mining [8]. Metaheuristics drawn from nature comprise the majority of nature-inspired algorithms. Two contrasting factors are frequently used when employing these algorithms: intensification (using the most effective solutions found) and diversity (investigation of the search area). Swarm algorithms have recently been proposed and demonstrated to perform admirably for FS problems. Most of the metaheuristic algorithms suffer from local optimum, lack of diversity of the search, and imbalance between the explorative or the exploitive capabilities of the algorithm [6].

Given the challenges that can hinder effective feature selection, we were motivated to develop a method that leverages the Grey Wolf Optimizer algorithm GWO. This approach addresses the issues associated with feature selection and aims to find more efficient solutions. The Grey Wolf Optimizer algorithm has demonstrated notable success across various applications, making it a strong candidate for this purpose.

To address the issues of local optima and limited diversity, the proposed algorithm incorporates a two-phase mutation operation (2M-GWO) to enhance its performance. This mutation operation is designed to improve the algorithm's exploitation capabilities, resulting in more effective and diverse solutions.

The following are the key contributions of our work:

- We propose a novel prediction model based on the 2M-GWO algorithm and five ML classifiers for SDP, which will select the most promising features and enhance the predictive capability of the prediction model. To the best of our knowledge, this model has never been employed for SDP.
- The performance of the proposed SDP model is assessed using 27 publicly available datasets to ensure the generalizability of the results. The investigation aimed to determine the impact of the improvements in 2M-GWO on optimization performance and robustness.
- A number of comparisons are made on the Receiver Operating Characteristic Area Under Curve (ROC_AUC) value using statistical concepts such as mean value and Friedman test. These methods offer a robust assessment of algorithm performance across different scenarios, providing insight into average ROC_AUC results and identifying significant variations among the compared algorithms.
- The proposed model based on the 2M-GWO algorithm is evaluated against five other metaheuristics techniques (Harris Hawks Optimization (HHO), Salp Swarm Optimization (SSO), Whale Optimization (WO), Jaya Optimization (JO), and Sine Cosine Optimization

(SCO)), and experimental results demonstrated its superiority over competing algorithms.

The organization of the research paper is as outlined: The review of the past literature is covered in Sect. 2, which discusses the various methodologies that have been used for SDP using evolutionary FS methods. In Sect. 3, the proposed SDP model based on 2M-GWO for FS is described. Section 4 covers the simulated results of the proposed model, along with qualitative and comparative analysis using statistical tools. Finally, Sect. 5 contains the conclusion and suggestions for future work.

2 Related work

The effectiveness of SDP models was examined in recent studies to determine how evolutionary FS techniques affected their performance. Table 1 presents a detailed description of existing literature related to the use of metaheuristic techniques in SDP.

3 Proposed methodology

To solve the SDP issue, we developed a wrapper strategy based on an enhanced version of the GWO algorithm. Figure 1 explores the mechanism of the suggested algorithm. A tenfold cross-validation (10-CV) procedure is performed at each iteration. Initially, the algorithm applies GWO on the SDP dataset. Using the 10-CV method, the chosen dataset is subdivided at each iteration. The SMOTE mechanism is used to resolve the unbalanced problem of the dataset on which the model should be trained. This procedure will generate a training dataset using oversampling. Five distinct classifiers (i.e., SVM, RF, GB, AB, and KNN) are used to construct an SDP model, which is then assessed using K-Fold = 10. The suggested method terminates whenever the best solution in terms of the objective function is attained, or the value of iterations reaches its maximum value. The model is shown in the following points.

3.1 Dataset

Defects are faults or errors that have been made in the past when developing software systems. These inaccuracies are used to forecast the likelihood of a fault or the total number of defects. Statistics on software defects can clearly show which software modules or components are particularly susceptible to faults. These undesirable defects can occur at many software systems levels and in a variety of granularities. Software metrics and developer information make up the majority of the data in software defect databases. Software attributes may be divided into process, product, and requirements attributes.

For this study's defect datasets, data from four different sources were collected (NASA, PROMISE, RELINK, and AEEEM). This study made use of the NASA dataset as it was created by Shepperd, et al. The NASA datasets include metrics generated from static code parts that take into account the size and complexity of the code [9]. Projects from the PROMISE repository are based on features that are gathered at the class level for object-oriented software and information about defects found in software components. This dataset was developed using the JAVA programming language and was retrieved from the Apache software [10]. The open-source RELINK dataset comprises three open-source projects (Apache HTTP, Safe, and Zxing) and has a total of twenty-six attributes [11].

AEEEM datasets differ from other data [12] as they include software attributes from source code measurements. The change metrics, entropy, and churn of source code metrics form the foundation of source code metrics. These datasets are obviously dependent on various software features. The datasets details are mentioned in Table 1.

3.2 Feature selection using 2M-GWO

3.2.1 Grey wolf algorithm (GWO)

Grey wolves (GWs) are superb predators with exceptional hunting abilities. This is due to the fact that they live in a disciplined and orderly pack. The GWO algorithm was recently presented to imitate the hunting, seeking, and surrounding behaviour of GWs [13]. There are four main varieties of GWs with varying levels of dominance and leadership in the wolf population (Table 2). The first, second, and third-best solutions in the GWO method are denoted by the letters α β and δ, correspondingly. The rest of the possible solutions are referred to as ω.

The hunting mechanism in GWO involves three main steps: encircling the prey, hunting (attacking), and searching for prey. These steps are mathematically modelled as follows:

Encircling Prey The wolves encircle the prey during the hunt. This behaviour is modelled using the position vectors of the prey and the wolves. The position of each wolf is updated according to the position of the prey using the following equations:

$$D_v = |C_v.X_v p - X_v(t)| \quad (1)$$

$$X_v(t+1) = X_v p(t) - A_v \cdot D_v \quad (2)$$

Table 1 Detailed description of existing related work

References	Dataset	Techniques	Benefits	Limitations
[30]	PROMISE Dataset: Ant (1.7), Camel (1.2, 1.4, 1.6), Jedit (3.2, 4.0 4.1, 4.2), Lucene (2.0, 2.2, 2.4), Log4j (1.0, 1.1, 1.2), Xalan (2.4, 2.5, 2.6)	Levy Sparrow Search (LSS), weighted dual cross recurrent network	The DCRN is used for its ability to handle complex data relationships, while the LSS algorithm helps in optimizing the model's parameters, enhancing the model's ability to predict faults accurately	The study only utilizes PROMISE datasets for validation. this could limit the understanding of the model's robustness and effectiveness across a broader range of applications
[31]	PROMISE Dataset (19 projects)	Mutation Boosted Salp Swarm Optimizer (MBSSO) Rough Set The ory, Kernel Extreme Learning Machine (KELM)	This study allows for effective feature selection based on relevance and dependency, leading to improved defect detection models with higher accuracy and efficiency	The study faces disadvantages such as complexity and high computational costs due to the integration of MBSSO with Rough Set Theory
[32]	PROMISE Dataset: KC (1, 3) PC (1, 2, 4, 5), MC1, JM1, MW1, CM1, MC2	Golden Jackal Optimization, K-Nearest Neighbor, Decision Tree, Quadrative Discriminant Analysis and Naive Bayes	The study presents a novel and effective feature selection method, FSGJO, that can significantly improve the performance of software fault prediction models by identifying the most relevant features	The study has limitations, such as limited generalizability, as the effectiveness of the FSGJO algorithm may be specific to the datasets used in the research
[24]	RELINK Dataset (3 projects), AEEEM Dataset (5 projects)	MIC_SM_FS (utilizing maximal information coefficient and statistical measures) and BPSO_FS (using binary particle swarm optimization), Naïve Bayes (NB), Random Forest (RF), and Logistic Regression (LR), TCA +	MIC_SM_FS selected features with high importance and similar distribution to the target project using maximal information coefficient and statistical measures. BPSO_FS optimized performance by selecting representative features through binary particle swarm optimization	If the source data is not representative or if it contains biases, the predictions made for the target projects might not be accurate
[33]	NASA Dataset: PC1, CM1, MC2, MW1 AEEEM Dataset: JDT, PDE, ML MORPH Dataset: ant-1.3, arc, camel-1.0, activemq-5.0., hbase-0.94.0, hive-0.9.0, groovy-1_6_BETA_1, jruby-1.1)	Sparrow Search algorithm, Extreme Learning Machine	The study presents a comprehensive approach to improving software defect prediction through a novel optimization algorithm and an ensemble learning framework, addressing key challenges such as parameter optimization, data imbalance, and model generalization	The scalability of the proposed algorithm in handling very large datasets is not discussed
[34]	PROMISE Dataset: Ant 1.7, Camel (1.0, 1.2, 1.4,1.6), Jedit (3.2, 4.0, 4.1, 4.2, 4.3), Log4j (1.0, 1.1), Lucene 2.0, Xalan (2.4, 2.5, 2.6)	Enhanced Binary Whale Optimization Algorithm, K-Nearest Neighbors (KNN), Naive Bayes (NB), Linear Discriminant Analysis (LDA), Linear Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), and RandomForest (RF)	The enhanced version of the BWOA referred to as improving the diversity and exploration capabilities of the BWOA, makes it more effective in navigating the feature selection search space	The generalization of these results to other types of software projects or datasets is not guaranteed

Table 1 (continued)

References	Dataset	Techniques	Benefits	Limitations
[35]	PROMISE Dataset: PC1, KC2, KC1, JM1, and CM1	War Strategy Optimization (WSO), Kernel-based Extreme Learning Machine (KELM)	The study reports an average increase of over 90% in accuracy across various parameters, which underscores the effectiveness of the combined approach in identifying defects in software components	The advanced algorithms used, such as WSO and KELM, may require substantial computational resources for training and optimization, especially when dealing with large datasets
[36]	PROMISE Dataset: Ar (1, 3, 4, 5, 6), Ant 1.7, CM1, Jedit (4.0, 4.1), Tomcat 6 NASA Dataset: JM1, MW1, KC (1, 2, 3), PC (1, 2, 4)	Mine Blast Algorithm and Simulated Annealing (MBA-SA), back propagation neural network (BPNN)	This hybrid approach aims to optimize the input connection weights and biases for a backpropagation neural network, significantly improving prediction accuracy across multiple datasets compared to existing methods	The suggested model may need significant computational resources for training and optimization
[37]	AEEEM Dataset: JDT, PDE, EQ	Bat Optimisation (BA), Improved Fisher's Linear Discriminant, Enhanced Deep Neural Network (EDNN)	This approach effectively reduces the dimensionality of the data, which in turn enhances the performance of the software bug prediction model. The IFLD method helps in selecting the most discriminative features from the software modules, which are crucial for accurate bug prediction	The study is limited to only the AEEEM dataset
[38]	PROMISE Dataset: Lucene NASA dataset: KC (1, 2), PC1, JM1	Genetic Programming, Adaboost	The proposed model tends to achieve superior predictive accuracy compared to simple statistical, ensemble, or individual machine-learning techniques alone	Due to the stochastic nature of genetic algorithms and the selection of features, there can be variability in the predictive performance
[39]	PROMISE Dataset: Camel (1.0, 1.2, 1.4, 1.6), Jedit (3.2, 4.0, 4.1,4.3), Ant-1.7, Log4j (1.1, 1.0, 1.2,), Xalan (2.4, 2.2)	Butterfly Optimization, Random Forest, ADASYN	The use of Butterfly optimization for feature selection in the context of Ensemble Random Forest with Adaptive synthetic sampling (E-RF-ADASYN) helps in identifying the most significant features for fault prediction	There is a risk of overfitting, where the model fits the training data too closely and fails to generalize well to unseen data
[40]	NASA Dataset: PC5, JM1, KC1	Binary Rao optimization, Naive Bayes (NB), Support Vector Machine (SVM)	The proposed approach in the study is a hybrid feature selection method that combines a filter-wrapper approach with a multi-criteria decision-making (MCDM) method and the Rao optimization algorithm. This method is designed to enhance the accuracy of software defect prediction models by selecting the most informative features from software metrics data	The study mentions the use of self-adaptive mechanisms to adjust parameters like population size, but the optimal configuration of these parameters can vary widely between different datasets and feature selection scenarios

Table 1 (continued)

References	Dataset	Techniques	Benefits	Limitations
[41]	NASA Dataset: CM1, JM1, KC1	Elitist Self-Adaptive Multi-Population Social Mimic Optimization (ESAMP-SMO)	This proposed model combines elements of social mimic optimization with genetic operations like crossover and mutation to enhance the search capabilities of the algorithm	The generalization of the method across diverse software projects with different characteristics and defect patterns is not explicitly addressed
[42]	PROMISE Dataset: Ant (1.6, 1.7), Camel (1.4, 1.6), Ivy 2.0, Jedit (4.2, 4.3), Poi 2.0, Prop 6, Synapse 1.2, Xalan (1.2, 2.5) RELINK Dataset: Apache, Safe, Zxing AEEEM Dataset: JDT, LC, EQ, PDE, ML	Whale Optimization Algorithm (WOA) and Simulated Annealing (SA), convolutional neural network (CNN) and kernel extreme learning machine (KELM	This introduced model effectively selects an optimal representative feature subset by leveraging the strong local search capability of SA to enhance the exploitation performance of WOA and the strong global search capability of WOA to boost the exploration of SA simultaneously	The computational complexity of combining WOA and SA could result in longer processing times, making the approach less efficient for large datasets
[43]	PROMISE Dataset: Ant 1.7, Camel (1.2, 1.4, 1.6), Jedit (3.2, 4.0, 4.1, 4.2), Log4j (1.0, 1.1, 1.2), Lucene (2.0, 2.2, 2.4), Xalan (2.4, 2.5, 2.6)	Boosted Whale Optimization Algorithm (BWOA), Decision Tree and K-Nearest Neighbour	The suggested improvement facilitates the WOA in avoiding local optima by augmenting its exploration process. Through thorough evaluation, it's evident that the proposed approach significantly enhances the performance of ML classification techniques, surpassing both the original WOA and six other commonly employed methods	The study is limited to the PROMISE dataset; the limitations of this study in terms of generalizability primarily revolve around the potential lack of diversity within the algorithm's application scope
[44]	PROMISE Dataset: Ant-1.7, Camel (1.0, 1.2, 1.4, 1.6), Jedit (3.2, 4.0, 4.1, 4.2), Log4j (1.0, 1.1, 1.2), Xalan (2.4, 2.7)	Enhanced Binary Harris Hawk Optimization (EBHHO), Selection K-nearest neighbours (kNN), Decision Trees (DT), and Linear Discriminant Analysis (LDA)	By reducing the dimensionality through feature selection, the algorithm improves the manageability and performance of the learning models used for predicting software faults	The use of only one benchmark dataset could limit the applicability of the findings to other types of datasets or real-world scenarios that have different characteristics
[29]	PROMISE Dataset: Ant 1.7, Camel (1.0, 1.2, 1.4, 1.6), Jedit (3.2, 4.0, 4.1, 4.2, 4.3), Log4j (1.0, 1.1, 1.2), Xalan-2.4, 2.7)	Enhanced Binary Moth Flame Optimization (EBMFO), k-nearest neighbours (k-NN), Decision Trees (DT), and Linear discriminant analysis (LDA)	The suggested approach improves the results from the literature and demonstrates the significance of transfer functions for FS methods	The proposed algorithm inherently carries the computational complexity for large datasets having more than 20 features

Table 1 (continued)

References	Dataset	Techniques	Benefits	Limitations
[45]	AEEEM Dataset: EQ, MORPH Dataset: Velocity 1.4, SOFTLAB Dataset: Ar1, NASA Dataset: CM1, PC2, RELINK Dataset: Apache	Firefly Particle Swarm Optimization, KNN, DT	The firefly algorithm helps in selecting the most influential features, while particle swarm optimization fine-tunes the prediction model	The integration of the firefly algorithm with particle swarm optimization is computationally intensive, which may limit its practical applicability in environments with limited computational resources or where rapid defect prediction is needed
[46]	NASA Dataset: CM1, PC1, KC1, KC2	Firefly Algorithm (FA) support vector machine (SVM), Naïve Bayes (NB) as well as K-nearest neighbour (KNN)	The Firefly Algorithm (FA) algorithm is particularly noted for its simplicity, efficiency, and ability to find global optima in multimodal optimization landscapes. The outcomes of the experiment indicate that the utilization of the FA in the FS can yield a classification efficiency that surpasses that of all other methods	The algorithm's effectiveness and efficiency might vary when applied to other datasets with different characteristics or higher dimensionality, which could affect its generalization capability
[47]	PROMISE Dataset: Ant-1.7, Camel (1.0, 1.2, 1.4, 1.6,) Jedit (3.4, 4.0, 4.2, 4.3), Log4j (1.0, 1.1, 1.2, 2.0, 2.2, 2.4), Xalan-2.4, 2.5, 2.6, 2.7)	Binary Genetic Algorithm (BGA), Binary Particle Swarm Optimization (BPSO), Binary Ant Colony Optimization (BACO), layered recurrent neural network (L-RNN),	The layered structure helps capture complex patterns and dependencies in the data, which improves the accuracy of software fault prediction. The findings demonstrate that the suggested strategy works comparatively better than other existing approaches	The study is limited to only PROMISE dataset
[48]	NASA Dataset: PC1, PC3, CM1, PROMISE Dataset: Ant-1.7, Ivy-2.0, Lucene 2.2, Tomcat, Xalan-2.4, Poi-2.0	Binary Teaching–Learning-Based Optimization (BTLBO), KNN	The approach produces a balanced selection of extremely important samples that may be used to train a classification technique for accurate prediction of possibly defective software products	The process of finding the optimal sample subset using BTLBO can be computationally expensive

$$A_v = 2a_v \cdot random_{v1} - a_v \tag{3}$$

$$C_v = 2 \cdot random_{v2} \tag{4}$$

$$a = 2 - t\left(\frac{2}{iterations}\right) \tag{5}$$

Here, $X_v(t)$ represents the position vector of a wolf at iteration t, $X_v p(t)$ represents the position of the prey. C_v and A_v are coefficient vectors. The parameter *iterations* determine the maximum number of iterations. The $random_{v1}$ and $random_{v2}$ are random vectors, whereas a_v decreases linearly from 2 to 0 over the course of iterations.

Attacking the Prey The alpha (best solution), beta, and delta wolves have better knowledge about the potential location of the prey. Therefore, the positions of the other wolves (omegas) are updated based on the positions of the alpha, beta, and delta wolves:

$$X_v(t+1) = \frac{X_{v1} + X_{v2} + X_{v3}}{3} \tag{6}$$

Here X_{v1}, X_{v2} and X_{v3} are the positions of the alpha, beta, and delta wolves, respectively, updated using similar equations to the encircling mechanism but cantered around each of these three wolves.

Searching for Prey To simulate the search for prey, the wolves diverge from each other and explore the search space. This is modelled by allowing the random vectors $random_{v1}$ and $random_{v2}$ to take values that can potentially increase the distance between the position of a wolf and the position of the prey.

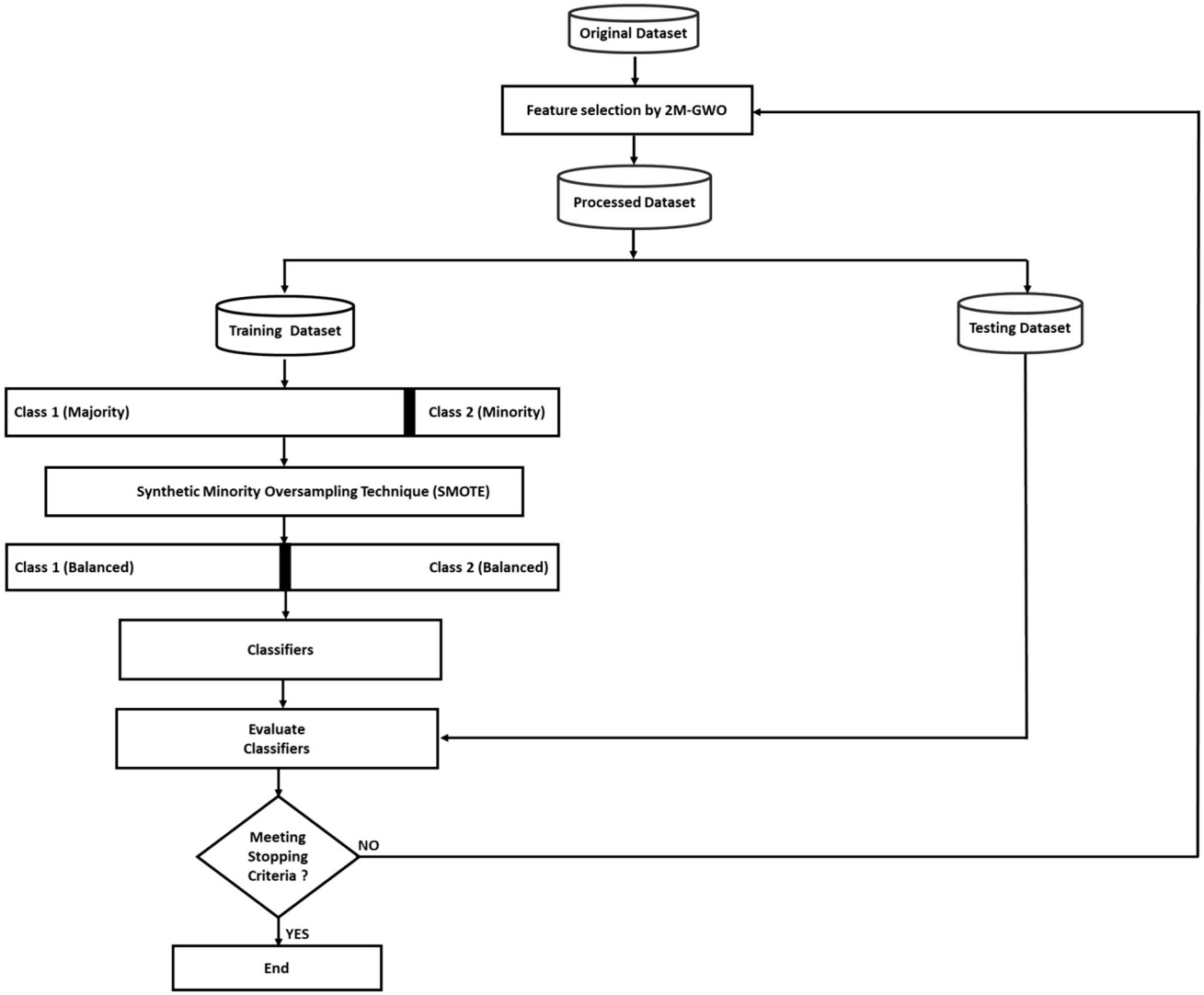


Fig. 1 Proposed methodology

Although the standard GWO is known for its effectiveness in various optimization tasks, but it faces challenges such as premature convergence and getting trapped in local optima when dealing with complex feature selection problems.

To address these challenges, the authors propose a two-phase mutation mechanism integrated into the GWO. This enhancement aims to improve the exploitation capabilities of the algorithm, making it more effective in navigating the search space of feature selection problems.

3.2.2 Two-phase mutation grey wolf optimisation (2M-GWO)

The Two-Phase Mutation Grey Wolf Optimizer (TMGWO) is an enhanced version of the standard GWO, specifically adapted for feature selection problems in classification tasks. This enhancement integrates a two-phase mutation mechanism to improve the exploitation capabilities of the algorithm, making it more effective in navigating the complex search space of feature selection problems. The benefits of using two mutation phases are as follows.

First Mutation Phase This phase aims to reduce the number of selected features while maintaining high classification accuracy. It selectively mutates certain features of the best solution (alpha wolf), turning some of the selected features off (from 1 to 0).

Second Mutation Phase This phase attempts to add potentially informative features that could increase classification accuracy. It operates by turning some unselected features on (from 0 to 1) based on their potential to improve the solution.

This 2M-GWO algorithm includes four steps: Initialization (step 1), evaluation (step 2), transformation

Table 2 Dataset details

Projects	#Instances	#Features	#Defective modules	% of Defective modules
PROMISE-Ant-1.7	745	20	166	22.28
PROMISE-Camel-1.2	608	20	216	35.5
PROMISE-Camel-1.4	872	20	145	16.6
PROMISE-Jedit-3.2	272	20	90	33.1
PROMISE-Jedit-4.1	312	20	79	25.32
PROMISE-Jedit-4.2	367	20	48	13.07
PROMISE-Log4j-1.0	135	20	34	25.18
PROMISE-Log4j-1.1	109	20	37	33.94
PROMISE-Lucene-2.0	195	20	91	46.66
PROMISE-Lucene-2.2	247	20	144	58.3
PROMISE-Tomcat	853	20	77	9.02
PROMISE-Xalan 2.4	723	20	110	15.2
PROMISE-Xalan 2.5	803	20	387	48.19
PROMISE-Xalan- 2.6	885	20	411	46.4
RELINK-Apache	194	26	98	50.52
RELINK-Safe	56	26	22	39.29
RELINK-Zxing	399	26	118	29.57
NASA-Cm1	328	37	42	12.8
NASA-Mw1	253	37	27	10.67
NASA-PC1	1109	37	77	6.94
NASA-PC3	1077	37	134	12.44
NASA-PC4	1458	37	178	12.21
AEEEM-EQ	325	61	129	39.69
AEEEM-JDT	997	61	206	20.66
AEEEM-LC	399	61	64	16.04
AEEEM-ML	1862	61	245	13.15
AEEEM-PDE	1492	61	209	14

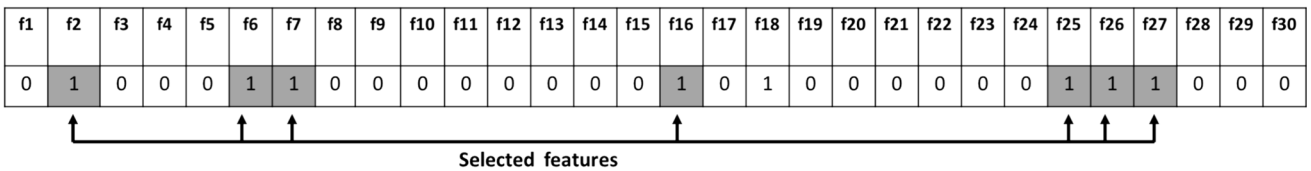


Fig. 2 Solution representation of 2M-GWO

function (step 3), and two-phase mutation (step 4) make up the core components of 2M-GWO.

Initialization The first step is to initialize a colony of "n" GWs or search agents (SA) generated randomly. Each SA represents a potential solution and has a size of "d" equal to the number of attributes in the actual dataset [14]. For classifying classes, the FS is used to identify the subset of features that produce the most accurate results. This is achieved by setting some features to "1" (selected) and others to "0" (not selected), as shown in Fig. 2. At the start; each solution is represented by binary values of either "0" or "1".

Evaluation The FS has multiple objectives since it must achieve two goals: reduce the count of features that must be selected and increase the classifier's accuracy. To achieve this balance, the fitness function is designed to evaluate the solutions and meet both objectives simultaneously in Eq. (7), where $\gamma_R(D)$ represents the rate of classification error of R, the condition attribute set in respect to decision D, which is computed via the ML classifier KNN.

$$fitness = \alpha\gamma_R(D) + \beta\frac{|s|}{|d|} \tag{7}$$

The span of the chosen attribute subset's cardinality is represented by the symbol |s|. Each sample of the actual dataset that contains features is represented by a cardinality of |d|. α and β are the two constraints of weight, which are equivalent to the value of the classification_accuracy and

Algorithm 1 The proposed 2M-GWO

Inputs: n, Number of grey wolves in the population.
iterations, maximum number of iterations.

1. Initialize a population of n grey wolves Xi, $i = 1$ to n
2. Initialize parameters a, A, and C
3. Calculate the fitness of each grey wolf
4. Assign the best three grey wolves to $X\alpha$, $X\beta$, and $X\delta$ respectively
5. Set $t = 0$

6. while $t <$ *iterations* do
   for each grey wolf Xi do
      Update the position of Xi using Eq. (6)
   end for
   Update parameters a, A, and C
   Convert the position of grey wolves to binary using Eq. (8) and Eq. (9)
   Calculate the fitness of all grey wolves
   Update the best three grey wolves $X\alpha$, $X\beta$, and $X\delta$
   Apply two-phase mutation to $X\alpha$ and update its fitness
   Increment t by 1
end while

return the best grey wolf $X\alpha$

Output: The best grey wolf $X\alpha$

the length of the subset of the chosen feature, $\alpha \in [0,1]$ and $\alpha + \beta = 1$. The evaluation function will overlook results that could have the equivalent accuracy but use fewer attributes, which is a key component in reducing the dimensionality problem if it solely considers classification accuracy.

Transformation function

The GW's positions produced by the typical GWO are continuous in nature. Since it conflicts with the binary structure of the FS, it cannot be as it is applied to our case. Since the issue with FS arises from picking (1) or ignoring (0) the advantageous attributes that improve the capability of the classifier. The task of changing the continuous value of the search space to binary values is performed by the transformation function (TF). The TF used in this study is an S-shaped sigmoidal function [15].

Conversion of binary values to continuous values is done by using the following Eqs. (8) and (9) of the sigmoidal transformation function:

$$x_i = \frac{1}{1 + e^{-x_i}} \tag{8}$$

$$x_{binary} = \begin{cases} 0 \; if \; rand < x_i \\ 1 \; if \; rand \geq x_i \end{cases} \tag{9}$$

where x_i depicts the continuous value of the feature, the value of i varies from 1 to d, x_{binary} can have a value of 0 (not selected) or 1 (selected) on comparing the x_i with a random value *rand.*

Two-phase mutation

A mutation operator is utilized to enhance the performance of the proposed algorithm. In the initial phase of mutation, the primary objective is to reduce the number of selected attributes while preserving high classification accuracy [13]. The subsequent phase focuses on improving classification efficacy by introducing more nuanced attributes. To mitigate delays that may arise from the mutation operator, step 4 can be executed with a mutation probability (Mp). This approach optimizes the mutation process, allowing for controlled adjustments without compromising efficiency.

Algorithm 1 provides the detailed pseudocode for the proposed mutation process. Additionally, the complete pseudocode for the 2M-GWO algorithm, incorporating both mutation and other key components, is outlined in Algorithm 2.

In Algorithm 2, $X\alpha$, $X\beta$, and $X\delta$ represent the positions of the best three grey wolves (alpha, beta, delta), which

Algorithm 2 Two-Phase Mutation

Input: Xα: the best grey wolf from each iteration

1. Calculate the fitness of Xα and store it as *fitness*
2. Initialize *one_positions* to store the indices of selected features in $X\alpha$
3. Set $Xmutated1 = X\alpha$

// First Mutation Phase
4. for each index *i* in *one_positions* do
   Generate a random number r
   if $r < Mp$ then
      Set *Xmutated*1 [*i*] = 0
      Calculate the fitness of *Xmutated*1 and store it as *fitness_mutated*
      if *fitness_mutated* < *fitness* then
         Set *fitness* = *fitness_mutated*
         Set $X\alpha$ = *Xmutated*1
      end if
   end if
end for

// Second Mutation Phase
5. Initialize *zero_positions* to store the indices of unselected features in $X\alpha$
6. Set $Xmutated2 = X\alpha$

7. for each index *j* in *zero_positions* do
   Generate a random number *r*
   if $r < Mp$ then
      Set *Xmutated*2 [*j*] = 1
      Calculate the fitness of *Xmutated*2 and store it as *fitness_mutated*
      if *fitness_mutated* < *fitness* then
         Set *fitness* = *fitness_mutated*
         Set $X\alpha$ = *Xmutated*2
      end if
   end if
end for
return $X\alpha$

Output: The improved $X\alpha$

guide the search process. *Xi* denotes the position of the i^{th} grey wolf. *Fitness* is the variable which shows the fitness value of *Xα*. *one_positions* is avector storing the positions of selected features in Xα, and *zero_positions* store the positions of unselected features in Xα.

Computational complexity of the 2M-GWO technique The overall computational complexity of the algorithm primarily pivots on the number of iterations t, the number of Grey Wolves n, the dimensionality d of each wolf, and the complexity f of the fitness evaluation function. The computational complexity expression, $O(t \times n \times (d + f))$, suggests that the algorithm scales linearly with the number of iterations and the number of wolves but also depends linearly on the sum of the dimensionality and the complexity of the fitness function. This indicates that while the algorithm can handle moderate sizes efficiently, large values or complex fitness functions could significantly increase computational demands.

3.3 Classifiers

The literature has a number of learning classifiers. As a result, we restrict our research to using only five distinct classifiers:

3.3.1 Support vector machine (SVM)

It aims to find the best hyperplane in a high-dimensional space that can successfully separate different classes. The hyperplane is defined by a set of parameters that are optimized during training to maximize the margin between the hyperplane and the closest data points from each class. One of the main advantages of SVM is its ability to handle high-dimensional data and small sa_{mp}le sizes [16].

3.3.2 Random forest (RF)

This is an example of an ensemble learning method used for classification, which involves creating multiple Decision Trees at training time and then combining the outputs of these trees to make a final decision. RF uses a bagging technique to train each DT on a random subset of the data used for training, which reduces overfitting and improves the generalizability of a model [17].

3.3.3 Gradient boosting (GB)

This is an ML technique that is commonly used for classifying software into defective and non-defective modules. It involves accumulating weak decision trees to a model on a recurring basis, with each new tree fitting the residual errors of the preceding tree. The model then makes a final decision by merging the outputs of all of the trees in the model. Gradient boosting can be used with a variety of loss functions, allowing it to handle a range of classification problems [18].

3.3.4 AdaBoost (AB)

The AB is also known as adaptive boosting, which trains a sequence of weak classifiers on the same dataset. Each subsequent classifier is trained to give more weight to the samples that were misclassified by the previous classifiers. Finally, the algorithm combines the predictions of all the weak classifiers to make a final prediction [10].

3.3.5 K-nearest neighbour (KNN)

The KNN works on the distance-based phenomenon. It is a non-parametric algorithm and is not dependent on the distribution of the given dataset. The KNN algorithm works by first training on a labelled dataset, and then using the learned information to classify new, unlabelled data points based on their resemblance to the training examples. The KNN algorithm determines the similarity between data points by calculating the distance among them, often using Euclidean distance [19].

3.4 Evaluation method

To assess the model's performance and facilitate result comparison, we employ ROC_AUC. It is a graphical tool that assesses a binary classifier's effectiveness by varying the discrimination threshold value. Ranging from 0 to 1, a score of 1 indicates a flawless classifier, while 0.5 signifies performance no better than random chance. The ROC_AUC curve illustrates the True Positive Rate (TPR) against the False Positive Rate (FPR) across different threshold values [20]. The True Positive Rate (TPR) is computed by dividing the number of correctly identified positive instances by the total number of positive instances, whereas the False Positive Rate (FPR) is determined by dividing the number of incorrectly identified positive instances by the total number of negative instances [21].

A high ROC_AUC value indicates that the classifier is able to distinguish between positive and negative cases well, while a low ROC_AUC value indicates poor performance. ROC_AUC is a popular evaluation metric other than accuracy in machine learning because it is insensitive to imbalances in class distribution and is well-suited for imbalanced datasets. Additionally, the ROC_AUC can be used to compare the performance of different classifiers.

3.5 Parameter setting

To address the stochastic character, each algorithm is performed 20 times independently using a random seed. The maximum number of iterations for all subsequent tests is set at 100. There are 5 SA in the entire population. Additionally, 10-CV is used in this case. In order to choose the optimal values for $\propto$ and β and based on certain values collected from past studies, several tests are carried out on various datasets. As a consequence, $\propto$ and β are set to corresponding values of 0.01 and 0.99. To enable a fair comparison of the algorithms, the parameter specifications of the techniques are collected from the existing studies. The parameters of some of the optimization techniques are mentioned in Table 3. For all other techniques, the default value is taken.

3.6 Statistical test

In software defect prediction, the Friedman test is often employed to evaluate and compare the effectiveness of various defect prediction models or techniques. This is crucial because software defect prediction models are designed to predict the likelihood of defects in software modules, and choosing the most effective model can significantly impact the quality and reliability of software.

Table 3 Parameter settings for optimization techniques

Technique	Parameter	Value
2M-GWO	Mutation probability, Mp	0.5
SCO	Constant value, a	2
WO	Spiral coefficient, b	1

The Friedman test serves as a non-parametric statistical method utilized for identifying variations among treatments across numerous test iterations. In the context of software defect prediction [22], it is particularly useful because it allows researchers to compare the performance of different prediction models or methods over multiple datasets without assuming a normal distribution of the data.

4 Result analysis

This subsection provides the findings in response to the Research Questions (RQ) as follows:

RQ 1 What is the effect of FS techniques on the SDP models? Does 2M-GWO significantly improve the performance of SDP models as compared to GWO and WFS Models?

Tables 4, 5, 6, 7, 8 and 9 explores the experimental results of the original GWO, modified 2M-GWO and WFS model for 27 datasets using five classifiers, namely SVM, RF, GB, AB and KNN. The experimental data include the feature size and ROC_AUC value for each dataset. We employed SMOTE to balance the classes in the given dataset. The ROC_AUC value for 27 open-source datasets is obtained by taking the average value of 20 runs for each project.

The data presented in Table 4 clearly demonstrates that the 2M-GWO algorithm outperforms the other two algorithms, showing promising results in terms of ROC_AUC values across a majority of projects. Analysis of average ROC_AUC values indicates that the modified 2M-GWO algorithm surpasses the original GWO and WFS methods, with a maximum average value of 0.785. Furthermore, the proposed algorithm, in conjunction with SVM, significantly improved average ROC_AUC values by over 6% across all datasets, demonstrating strong overall performance.

Table 5 illustrates the results of the RF classifier utilized to evaluate the original GWO, the modified 2M-GWO, and the complete dataset. It is evident that the 2M-GWO algorithm outperforms the other two methods, as it demonstrates the highest ROC_AUC values across a larger number of projects. The enhanced 2M-GWO surpasses the original GWO and WFS techniques in terms of average ROC_AUC values, with a maximum average value of 0.8. The average ROC_AUC value for all datasets saw an increase of approximately 2% when utilizing the proposed algorithm.

Table 4 ROC_AUC by SVM

Projects	GWO	2M-GWO	WFS
PROMISE-Ant-1.7	0.821	0.815	**0.88**
PROMISE-Camel-1.2	0.515	**0.618**	**0.618**
PROMISE-Camel-1.4	**0.704**	0.629	0.629
PROMISE-Jedit-3.2	0.779	**0.81**	0.782
PROMISE-Jedit-4.1	**0.87**	0.783	0.725
PROMISE-Jedit-4.2	0.747	**0.83**	0.725
PROMISE-Log4j-1.0	0.912	**0.914**	0.9
PROMISE-Log4j-1.1	0.867	0.88	**1**
PROMISE-Lucene-2.0	0.765	**0.807**	**0.807**
PROMISE-Lucene-2.2	**0.656**	0.589	0.589
PROMISE-Tomcat	0.709	**0.737**	**0.737**
PROMISE-Xalan 2.4	**0.757**	0.672	0.672
PROMISE-Xalan 2.5	0.603	**0.752**	0.647
PROMISE-Xalan-2.6	0.649	**0.961**	0.752
AEEEM-EQ	0.801	0.803	**0.825**
AEEEM-JDT	0.836	**0.848**	0.556
AEEEM-LC	0.801	**0.868**	0.556
AEEEM-ML	0.772	**0.798**	0.556
AEEEM-PDE	0.653	**0.755**	0.667
NASA-Cm1	0.751	**0.795**	0.698
NASA-Mw1	0.578	**0.777**	0.683
NASA-PC1	0.736	0.846	**0.849**
NASA-PC3	0.812	**0.896**	0.852
NASA-PC4	0.851	0.906	**0.919**
RELINK-Apache	0.732	**0.776**	0.689
RELINK-Safe	0.629	0.629	**0.886**
RELINK-Zxing	0.667	**0.714**	0.705
Avg	0.739	0.785	0.737

The significance of bold indicates the highest value for each dataset when performing the comparative analysis

The ROC_AUC value calculated by Gradient Boosting (GB) for the full dataset and processed datasets, in which features were selected GWO and 2M-GWO, is presented in Table 6. The results indicate that the 2M-GWO-based model demonstrates superior performance compared to the GWO and WFS cases, with an average improvement of 2.3% in the ROC_AUC value. In approximately 40% of datasets, the 2M-GWO-based model outperformed the others, while in approximately 19% of datasets, its performance was comparable to either the GWO-based model or the full dataset.

Table 7 presents the computed ROC_AUC value by AB for the entire datasets and the processed datasets, where the

Table 5 ROC_AUC by RF

Projects	GWO	2M-GWO	WFS
PROMISE-Ant-1.7	0.832	**0.845**	0.843
PROMISE-Camel-1.2	0.641	0.634	**0.652**
PROMISE-Camel-1.4	**0.736**	0.71	0.715
PROMISE-Jedit-3.2	0.775	**0.776**	0.758
PROMISE-Jedit-4.1	**0.886**	0.862	0.692
PROMISE-Jedit-4.2	**0.733**	0.73	0.692
PROMISE-Log4j-1.0	0.879	**0.889**	0.871
PROMISE-Log4j-1.1	0.876	**1**	**1**
PROMISE-Lucene-2.0	0.73	**0.767**	0.749
PROMISE-Lucene-2.2	0.563	**0.616**	**0.616**
PROMISE-Tomcat	0.811	**0.83**	0.824
PROMISE-Xalan 2.4	**0.754**	0.749	0.715
PROMISE-Xalan 2.5	0.656	**0.827**	0.748
PROMISE-Xalan- 2.6	0.82	**1**	0.814
AEEEM-EQ	0.818	**0.841**	0.825
AEEEM-JDT	0.877	**0.881**	0.85
AEEEM-LC	**0.884**	0.821	0.849
AEEEM-ML	**0.859**	0.838	0.857
AEEEM-PDE	0.787	0.784	**0.793**
NASA-Cm1	0.623	0.718	**0.766**
NASA-Mw1	**0.733**	0.5	0.613
NASA-PC1	0.927	0.925	**0.93**
NASA-PC3	0.825	**0.835**	0.827
NASA-PC4	0.919	0.931	**0.938**
RELINK-Apache	**0.732**	0.697	0.68
RELINK-Safe	0.786	0.786	**0.971**
RELINK-Zxing	0.611	**0.827**	0.807
Avg	0.780	**0.800**	0.792

The significance of bold indicates the highest value for each dataset when performing the comparative analysis

Table 6 ROC_AUC by GB

Projects	GWO	2M-GWO	WFS
PROMISE-Ant-1.7	**0.78**	0.71	0.71
PROMISE-Camel-1.2	0.514	**0.551**	**0.551**
PROMISE-Camel-1.4	0.726	**0.673**	**0.673**
PROMISE-Jedit-3.2	**0.823**	0.771	0.771
PROMISE-Jedit-4.1	0.786	**0.819**	0.702
PROMISE-Jedit-4.2	0.687	**0.731**	0.702
PROMISE-Log4j-1.0	0.7	**0.8**	0.764
PROMISE-Log4j-1.1	0.752	**0.933**	0.933
PROMISE-Lucene-2.0	**0.754**	0.728	0.728
PROMISE-Lucene-2.2	**0.714**	**0.714**	**0.714**
PROMISE-Tomcat	0.758	**0.805**	**0.805**
PROMISE-Xalan 2.4	**0.731**	0.72	0.72
PROMISE-Xalan 2.5	0.663	0.656	**0.666**
PROMISE-Xalan- 2.6	0.798	**0.983**	0.797
AEEEM-EQ	0.836	**0.841**	0.828
AEEEM-JDT	0.78	**0.82**	0.784
AEEEM-LC	**0.868**	0.803	0.764
AEEEM-ML	0.787	**0.805**	0.804
AEEEM-PDE	0.737	0.697	**0.748**
NASA-Cm1	0.519	**0.656**	0.588
NASA-Mw1	**0.613**	0.561	0.535
NASA-PC1	0.9	0.895	**0.907**
NASA-PC3	**0.81**	0.798	0.801
NASA-PC4	0.911	0.926	**0.931**
RELINK-Apache	0.683	**0.789**	0.684
RELINK-Safe	**0.857**	**0.857**	0.7
RELINK-Zxing	0.738	**0.747**	0.717
Avg	0.749	**0.767**	0.741

The significance of bold indicates the highest value for each dataset when performing the comparative analysis

features were selected by GWO and 2M-GWO. It is observed that the 2M-GWO-based model outperforms the GWO and WFS case in terms of overall performance, with an average ROC_AUC value improvement of 4.6%. Specifically, for approximately 33% of the datasets, the 2M-GWO-based model performs better, while for around 11% of the datasets, it performs similarly to either of the two approaches.

According to the data presented in Table 8, the 2M-GWO-KNN demonstrates superior classification performance compared to the other model in over 50% of the datasets, as indicated by the ROC_AUC value. The average ROC_AUC value for the overall performance of 2M-GWO shows an increase of approximately 4% when compared to the other two models.

Based on the analysis presented in Table 9 and Fig. 3, on taking an average value, the 2M-GWO algorithm achieved an approximate 80% reduction in feature sets across various datasets. The NASA dataset showed the most significant reduction at around 86%, followed by the AEEEM and RELINK datasets, which saw an 80% reduction each. In contrast, the PROMISE dataset had the smallest reduction in features at 75%.

Table 10 presents the comparative performance of different feature selection techniques based on their mean rank achieved by Friedman's test, determined through the ROC_AUC metric. The mean rank serves as a measure of each technique's efficacy in enhancing the prediction accuracy of a software defect prediction model. A lower mean rank indicates better performance. The p-value associated with these rankings is 0.022, suggesting that the differences in performance among these techniques are statistically significant at the 95% confidence interval. 2M-GWO stands out as the most effective technique, achieving the highest performance with a mean rank of 1.00, denoted as rank I. GWO is ranked second with a mean rank of 2.40,

Table 7 ROC_AUC by AB

Projects	GWO	2M-GWO	WFS
PROMISE-Ant-1.7	0.741	**0.793**	0.693
PROMISE-Camel-1.2	0.507	**0.517**	0.517
PROMISE-Camel-1.4	**0.732**	0.714	0.714
PROMISE-Jedit-3.2	**0.791**	0.78	0.78
PROMISE-Jedit-4.1	0.789	**0.822**	0.714
PROMISE-Jedit-4.2	0.673	0.692	**0.714**
PROMISE-Log4j-1.0	0.757	**0.793**	0.657
PROMISE-Log4j-1.1	0.752	**0.933**	**0.933**
PROMISE-Lucene-2.0	**0.77**	0.767	0.767
PROMISE-Lucene-2.2	0.573	**0.704**	**0.704**
PROMISE-Tomcat	0.659	**0.731**	0.721
PROMISE-Xalan 2.4	**0.738**	0.734	0.724
PROMISE-Xalan 2.5	0.68	**0.806**	0.648
PROMISE-Xalan- 2.6	0.807	**1**	0.806
AEEEM-EQ	0.766	**0.845**	0.818
AEEEM-JDT	0.819	0.825	**0.831**
AEEEM-LC	**0.845**	0.778	0.577
AEEEM-ML	0.787	0.786	**0.803**
AEEEM-PDE	0.673	0.643	**0.702**
NASA-Cm1	0.588	0.603	**0.642**
NASA-Mw1	0.57	0.6	**0.613**
NASA-PC1	0.909	0.899	**0.916**
NASA-PC3	**0.796**	0.768	0.705
NASA-PC4	0.91	0.92	**0.924**
RELINK-Apache	0.732	**0.775**	0.589
RELINK-Safe	**0.914**	**0.914**	0.729
RELINK-Zxing	0.601	0.715	**0.719**
Avg	0.736	**0.772**	0.728

The significance of bold indicates the highest value for each dataset when performing the comparative analysis

Table 8 ROC_AUC value calculated by KNN

Projects	GWO	2M-GWO	WFS
PROMISE-Ant-1.7	0.741	0.806	**0.815**
PROMISE-Camel-1.2	0.669	**0.693**	0.643
PROMISE-Camel-1.4	**0.699**	0.686	0.686
PROMISE-Jedit-3.2	0.733	**0.84**	**0.84**
PROMISE-Jedit-4.1	**0.874**	0.855	0.638
PROMISE-Jedit-4.2	**0.793**	0.654	0.638
PROMISE-Log4j-1.0	0.821	**0.868**	0.829
PROMISE-Log4j-1.1	0.943	**0.948**	**0.948**
PROMISE-Lucene-2.0	0.783	0.788	**0.79**
PROMISE-Lucene-2.2	**0.643**	0.624	0.584
PROMISE-Tomcat	0.628	**0.647**	**0.647**
PROMISE-Xalan 2.4	0.719	**0.79**	**0.79**
PROMISE-Xalan 2.5	0.629	**0.81**	0.665
PROMISE-Xalan- 2.6	0.814	**1**	0.81
AEEEM-EQ	0.782	0.84	**0.871**
AEEEM-JDT	**0.842**	0.78	0.763
AEEEM-LC	0.848	**0.851**	0.763
AEEEM-ML	0.785	0.786	**0.811**
AEEEM-PDE	0.657	**0.783**	0.715
NASA-Cm1	0.651	0.706	**0.759**
NASA-Mw1	0.596	0.524	**0.624**
NASA-PC1	0.905	**0.941**	0.852
NASA-PC3	0.78	0.791	**0.833**
NASA-PC4	0.852	**0.905**	0.827
RELINK-Apache	0.683	**0.689**	0.671
RELINK-Safe	0.814	**0.814**	0.786
RELINK-Zxing	0.67	0.782	**0.793**
Avg	0.753	**0.785**	0.755

The significance of bold indicates the highest value for each dataset when performing the comparative analysis

labelled as rank II. This shows that the standard Grey Wolf Optimiser, while effective, is outperformed by its two-phase variant in the context of feature selection for defect prediction. WFS is closely ranked third with a mean rank of 2.60, marked as rank III.

Summary: RQ1 reveals that it is apparent that the 2M-GWO method is an effective approach for reducing feature size while maintaining model performance for SDP for all five datasets. The 2M-GWO technique emerges as the most effective approach, achieving the highest level of performance with a mean rank of 1.00. Additionally, the 2M-GWO algorithm facilitated an average reduction of approximately 80% in feature sets across multiple datasets

RQ 2 Which is the best-performing ML technique for predicting software defects?

Table 11 displays the mean ranks obtained by the Friedman's test on the basis of ROC_AUC for the ML techniques used in the study. 2M-GWO-RF achieved the best mean rank of 2.19, indicating its superior performance in terms of prediction accuracy and efficiency. 2M-GWO-KNN attained a mean rank of 2.94, placing it second in terms of performance. While it is not as effective as 2M-GWO-RF, it still demonstrates good performance and is viable with a mean rank of 3.15; SVM, along with 2MGWO, is ranked third in terms of performance. AB and GB are the ensemble classifiers with a mean rank of 3.35 and 3.37, ranking fourth and last among the classifiers evaluated. The p-value of 0.034 indicates that the differences in the mean ranks of these techniques are statistically significant. Overall, the combination of 2M-GWO and

Table 9 Number of features calculated by different techniques

Projects	GWO	2M-GWO	WFS
PROMISE-Ant-1.7	**5**	6	20
PROMISE-Camel-1.2	**6**	**6**	20
PROMISE-Camel-1.4	**5**	6	20
PROMISE-Jedit-3.2	**5**	**5**	20
PROMISE-Jedit-4.1	5	**4**	20
PROMISE-Jedit-4.2	**4**	**4**	20
PROMISE-Log4j-1.0	**4**	**4**	20
PROMISE-Log4j-1.1	**3**	**3**	20
PROMISE-Lucene-2.0	6	**5**	20
PROMISE-Lucene-2.2	7	**6**	20
PROMISE-Tomcat	**4**	**4**	20
PROMISE-Xalan 2.4	6	**5**	20
PROMISE-Xalan 2.5	10	**7**	20
PROMISE-Xalan- 2.6	7	**6**	20
AEEEM-EQ	**11**	12	61
AEEEM-JDT	14	**13**	61
AEEEM-LC	**6**	**6**	61
AEEEM-ML	**12**	14	61
AEEEM-PDE	**12**	13	61
NASA-Cm1	4	**2**	37
NASA-Mw1	**4**	**4**	37
NASA-PC1	**4**	**4**	37
NASA-PC3	8	**6**	37
NASA-PC4	**8**	**8**	37
RELINK-Apache	**5**	6	26
RELINK-Safe	3	3	26
RELINK-Zxing	8	**7**	26
Avg	7	**6**	31

The significance of bold indicates the highest value for each dataset when performing the comparative analysis

Random Forest (2M-GWO-RF) appears to offer the best performance among the tested classification techniques.

Summary: From RQ 2, it is evident that the combination of 2M-GWO and Random Forest (2M-GWO-RF) is the best-performing machine-learning technique for predicting software defects, as indicated by its superior mean rank of 2.19

RQ 3 Which software metrics are frequently selected using the proposed algorithm?

The analysis of feature selection across various datasets using the 2M-GWO FS technique is detailed in Figs. 4, 5, 6 and 7. These figures graphically represent the selection frequency of specific software metrics (features) crucial for predicting software defects. The x-axis of each figure indicates different software metrics derived from the

Table 10 Mean Rank obtained on the basis of ROC_AUC for feature selection technique

Techniques	Mean rank (p value = 0.022)
2M-GWO	1.00 (I)
GWO	2.40 (II)
WFS	2.60 (III)

Table 11 Mean Rank obtained on the basis of ROC_AUC for ML techniques

Classification Techniques	Mean rank (p value = .034)
2M-GWO-RF	2.19 (I)
2M-GWO-KNN	2.94 (II)
2M-GWO-SVM	3.15 (III)
2M-GWO-AB	3.35 (IV)
2M-GWO-GB	3.37 (V)

datasets, while the y-axis quantifies how frequently each metric was selected across multiple iterations.

For the RELINK dataset (Fig. 4), metrics such as CountLineCodeDecl, CountLineBlank, and CountLineCode show high selection frequencies of 26, 21, and 20 times, respectively, out of 60 total runs distributed across three projects (each executed 20 times). In the case of the PROMISE dataset (Fig. 5), the object-oriented metrics rfc, ce, and npm were selected 131 and 102 times each out of 280 iterations. Regarding the AEEEM dataset (Fig. 6), which includes 61 process metrics across five projects, metrics such as CVSEntropy, numberOfNontrivialBugsFoundUntil, and Ck OO numberOfMethodsInherited were most frequently selected, with counts of 49, 45, and 41 respectively, out of 100 runs. For the NASA project (Fig. 7), which focuses on Halstead's and size metrics, the metrics LOC CODE AND COMMENT, LOC BLANK, and LOC EXECUTABLE were identified as the most frequently occurring, with each being selected 28, 25, and 25 times respectively in twenty runs across five projects.

Summary: Analysis from RQ3 reveals key patterns in metric selection across different datasets. For the RELINK dataset, metrics such as CountLineCodeDecl, CountLineBlank, and CountLineCode were consistently chosen. In the PROMISE dataset, the most frequently selected metrics were object-oriented measures, including rfc, ce, and npm. In the AEEEM dataset, encompassing 61 process metrics across five projects, CVSEntropy, numberOfNontrivialBugsFoundUntil, and Ck OO numberOfMethodsInherited were the metrics most often chosen. In the NASA project, which focuses on Halstead's size metrics, LOC CODE AND COMMENT, LOC BLANK, and LOC EXECUTABLE were the most prominent metrics. The frequent selection of these metrics underscores their importance in the defect prediction process, suggesting that they provide valuable insights into software quality and should be considered in data collection for future projects

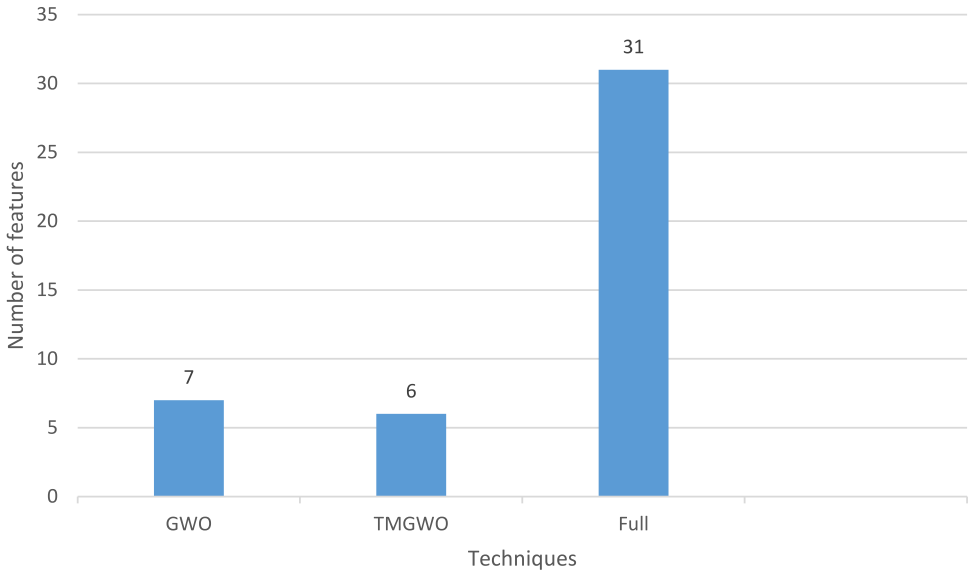


Fig. 3 Comparison of feature selection techniques on the basis of the number of features selected

RQ 4 What is the predictive performance of the proposed model as compared to other benchmark FS models for SDP?

In order to conduct a comprehensive comparative analysis, two types of analysis are made.

In first type, the proposed algorithm is evaluated against five closely related meta-heuristic algorithms: Harris Hawks Optimization (HHO) [23], Salp Swarm Optimization (SSO) [24], Whale Optimization (WO) [25], Jaya Optimization (JO) [26], and Sine Cosine Optimization (SCO) [27]. These comparisons, summarized in Table 12, were carried out under consistent experimental conditions to ensure fairness and accuracy in the results. Each algorithm was implemented and executed within the same computational environment, adhering to identical parameter settings. This approach facilitated direct comparisons and minimized variations that could arise from differing configurations. The ROC_AUC measure was used as the primary basis for evaluating performance across all algorithms, as it provides a comprehensive measure of an algorithm's ability to distinguish between classes or optimize solutions effectively. By examining the AUC scores, we analyzed the fact that the proposed technique outperforms the other techniques in approximately 55.5% of the total datasets. SSO obtained a promising outcome in approximately 18.5% of the datasets. HHO and WO resulted in the best ROC_AUC value for two datasets, whereas SCO and JO outperformed for only one dataset each. This showcases the effectiveness of the proposed method in selecting the most pertinent features, which notably boost the performance of the SDP model.

In the second type, the performance outcomes of established approaches are sourced from existing literature and subjected to a thorough comparison. In the RKEE-based model, Riaz et al. [28] proposed two-stage data preprocessing which incorporates FS and a new rough set Easy Ensemble scheme, which shows an average value of 0.771, whereas the existing model proposed by Tumar et al. [29] was based on improved binary Moth Flame Optimization with Adaptive synthetic sampling (ADASYN) to predict software defects. BMFO is employed as a wrapper FS, while ADASYN enhances the input dataset, addresses the imbalanced dataset and shows a ROC_AUC value of 0.7552. When compared with the existing methods based on evolutionary FS techniques, the proposed model based on 2M-GWO and Random Forest classifier obtained a ROC_AUC of 0.800, which is better than the existing methods.

Summary: Research Question 4 demonstrates that the proposed model, integrating 2M-GWO with Random Forest, surpasses the performance of five related metaheuristic techniques in more than 55% of the datasets. Furthermore, it exhibits competitive results when evaluated against current state-of-the-art approaches

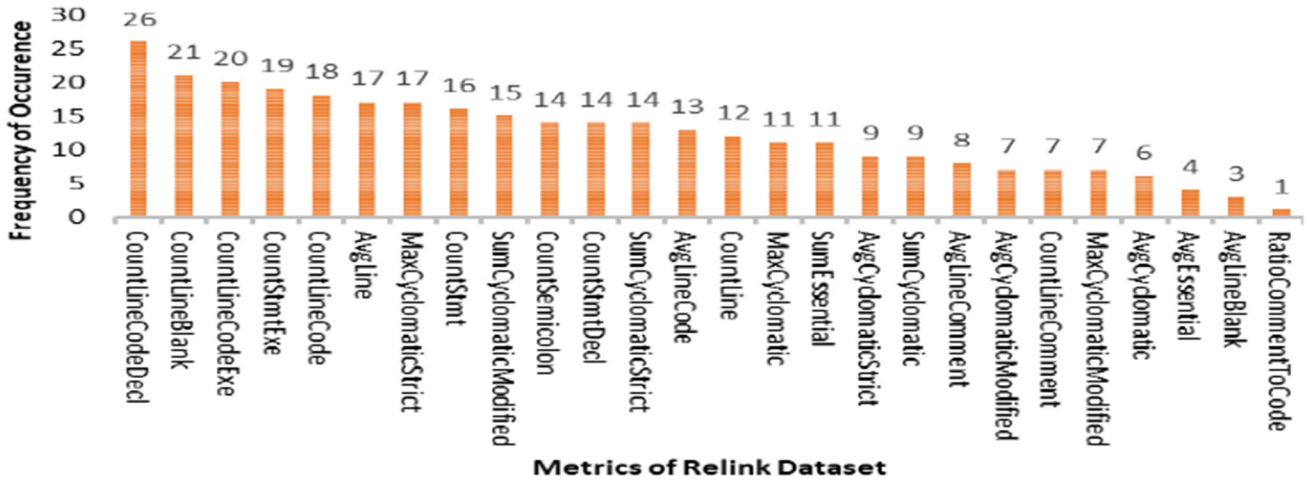


Fig. 4 Occurrence of metrics for RELINK dataset

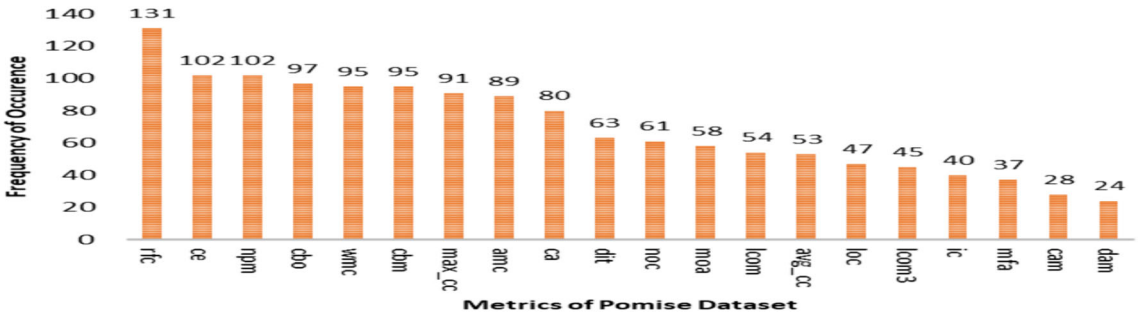


Fig. 5 Occurrence of metrics for PROMISE dataset

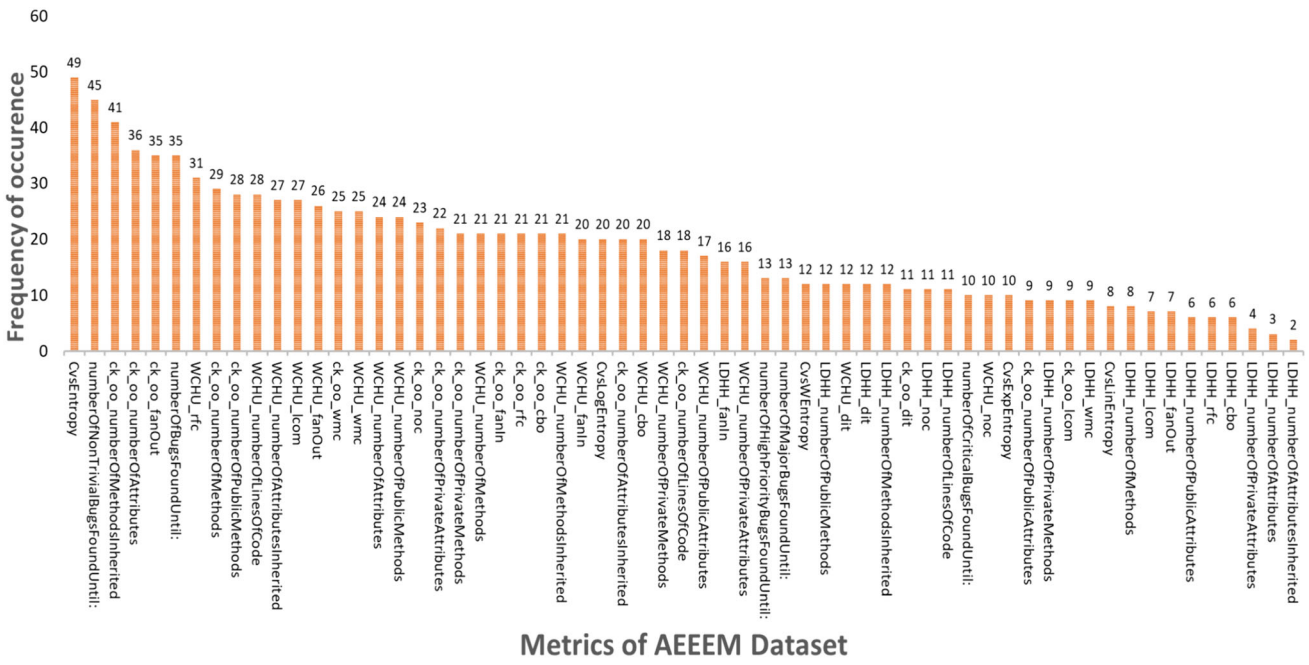


Fig. 6 Occurrence of metrics for AEEEM dataset

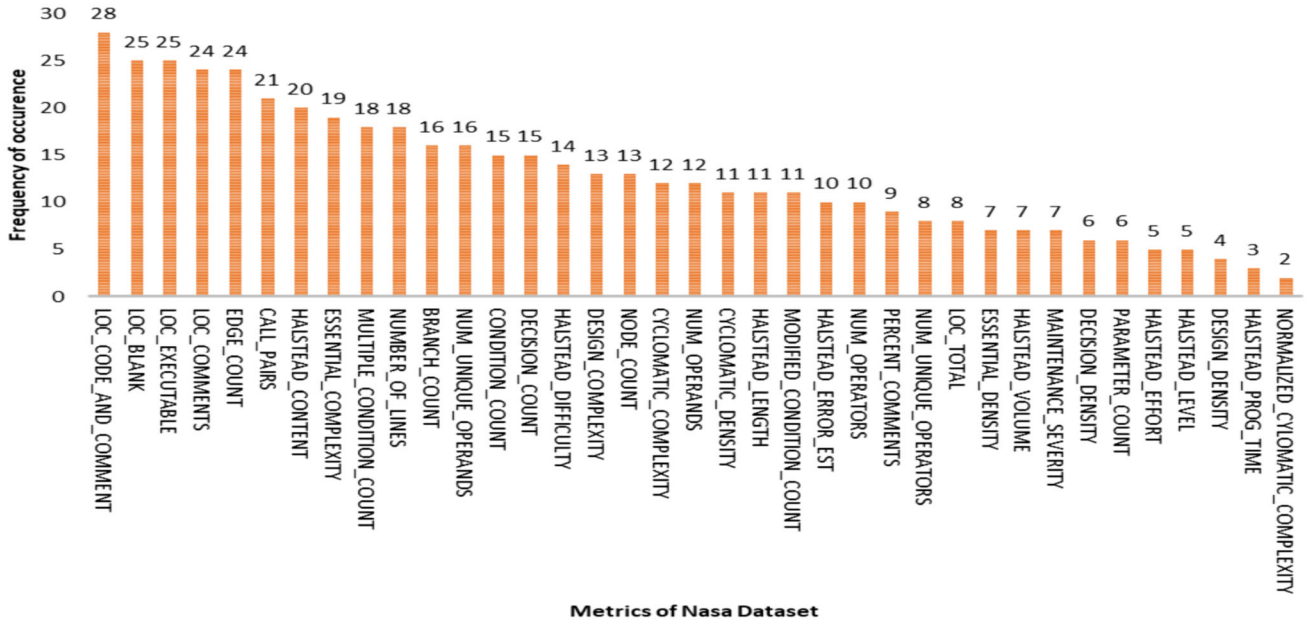


Fig. 7 Occurrence of metrics for NASA dataset

Table 12 ROC_AUC values of Proposed and other metaheuristic techniques

Projects	2 M-GWO -RF	SCO-RF	SSO-RF	JO-RF	WO-RF	HHO-RF
PROMISE-Ant-1.7	**0.845**	0.824	0.83	0.829	0.82	0.818
PROMISE-Camel-1.2	**0.634**	0.625	0.604	0.615	0.63	0.615
PROMISE-Camel-1.4	0.71	0.787	0.729	0.753	0.765	**0.798**
PROMISE-Jedit-3.2	0.776	0.799	0.799	0.808	0.784	**0.831**
PROMISE-Jedit-4.1	**0.862**	0.807	0.842	0.813	0.809	0.85
PROMISE-Jedit-4.2	**0.73**	0.65	0.615	0.685	0.712	0.727
PROMISE-Log4j-1.0	**0.889**	0.875	0.8174	0.865	0.812	0.801
PROMISE-Log4j-1.1	**1**	0.954	**1**	0.957	0.964	0.997
PROMISE-Lucene-2.0	0.767	0.767	0.76	0.769	**0.789**	0.785
PROMISE-Lucene-2.2	0.616	0.645	**0.664**	0.614	0.623	0.6125
PROMISE-Tomcat	**0.83**	0.799	0.802	0.772	0.825	0.809
PROMISE-Xalan 2.4	0.749	0.804	**0.849**	0.816	0.7182	0.788
PROMISE-Xalan 2.5	0.827	0.824	0.824	0.823	**0.877**	0.812
PROMISE-Xalan-2.6	**1**	0.988	0.998	0.991	0.962	0.99
AEEEM-EQ	0.841	**0.897**	0.839	0.896	0.875	0.807
AEEEM-JDT	**0.881**	0.814	0.881	0.805	0.803	0.8169
AEEEM-LC	**0.821**	0.82	0.803	0.805	0.8033	0.805
AEEEM-ML	**0.838**	0.822	0.837	0.801	0.811	0.808
AEEEM-PDE	0.784	0.775	0.709	0.773	**0.823**	0.807
NASA-Cm1	**0.718**	0.708	0.682	0.705	0.713	0.674
NASA-Mw1	0.5	0.655	0.604	0.615	0.68	**0.691**
NASA-PC1	**0.925**	0.888	0.894	0.905	0.893	0.817
NASA-PC3	0.835	0.87	**0.886**	0.884	0.872	0.8176
NASA-PC4	**0.931**	0.917	0.902	0.879	0.8919	0.852
RELINK-Apache	**0.697**	0.655	0.604	0.615	0.618	0.691
RELINK-Safe	0.786	0.737	**0.817**	0.785	0.802	0.816
RELINK-Zxing	0.827	0.773	0.813	**0.886**	0.836	0.814

The significance of bold indicates the highest value for each dataset when performing the comparative analysis

5 Threats to validity

This section explores the potential threats to the validity of the study and outlines its limitations.

5.1 Internal validity

To ensure internal validity, we have used a modified version of the GWO algorithm, which incorporates a two-phase mutation operation to improve the algorithm's exploitation capability. We have used a sigmoid function to transform the continuous search space to a binary one, matching the binary nature of feature selection. We have also employed a two-phase mutation to reduce the number of selected features while preserving high classification accuracy.

5.2 External validity

The external validity pertains to the degree to which the findings of an empirical study can be generalized and applied beyond the specific conditions of the study. In this research, the proposed algorithm's performance was assessed across 27 datasets. This wide range of datasets ensures the generalizability of the results to various types of datasets and classification problems.

5.3 Construct validity

Construct validity ensures that both independent and dependent variables accurately represent the intended concepts. This paper does not encounter any threats to construct validity, as the predictors utilized in the study are commonly employed in predictive modelling within the domain of software engineering.

5.4 Conclusion validity

This concerns the degree to which the conclusions drawn in the study are robust. Statistical analyses were conducted to validate the efficacy of the proposed algorithm and its superiority. This encompassed comparative assessments between the proposed algorithm and other metaheuristic algorithms, with the results subjected to statistical scrutiny employing the Friedman Test and Wilcoxon Signed Rank Test.

The study is limited to the use of public/open-source datasets, which can lead to constraints in terms of the generalizability of the SDP model's findings to a broader range of software applications, particularly those in the industrial sector. Open-source projects typically exhibit notable differences from proprietary or industrial projects in terms of development methodologies, team compositions, and code standards. These differences can impact the effectiveness of our proposed work when applied to industrial/proprietary software projects. In future, we aim to use the proposed SDP model for proprietary datasets from several industries to evaluate its predictive performance and improve the model for wider applicability.

6 Conclusion and future work

In this study, we proposed a prediction model based on 2M-GWO for SDP, which incorporates FS to enhance the quality of the ML techniques, i.e., SVM, AB, GB, NB and KNN, are used for the classification of the binary class dataset. We systematically designed the experiments based on 27 datasets from AEEEM, PROMISE, RELINK and NASA datasets. The key findings of the study are as follows:

- The proposed two-phase Grey Wolf Optimizer-based feature selection method, paired with the Random Forest classifier (2M-GWO-RF), effectively enhances software defect prediction (SDP). By reducing feature size, the 2M-GWO method improves model performance across multiple datasets and supports the accurate identification and removal of redundant and irrelevant features.
- The 2M-GWO-RF model achieves superior results, ranking highest in terms of ROC_AUC with a mean rank of 2.19, demonstrating exceptional prediction accuracy and efficiency (RQ2).
- Statistical analysis using Friedman's test confirms the significance of the mean rank differences ($p = 0.034$), validating 2M-GWO-RF as the top classification technique.
- The proposed model's robustness is evident through its focus on key metrics across various datasets, including CountLineCodeDecl, rfc, ce, npm, CVSEntropy, and LOC CODE AND COMMENT, among others, indicating its potential for future data collection and analysis.
- The 2M-GWO-RF model outperforms five related metaheuristic techniques in over 55% of the datasets, showcasing its superior choice in comparison to state-of-the-art approaches.

The study concludes that the two-phase Grey Wolf Optimizer with Random Forest classifier (2M-GWO-RF) significantly enhances software defect prediction by

offering superior feature selection and high prediction capability.

Acknowledgements Not applicable.

Author contributions RM and KK jointly contributed to the conceptualization and design of the study. KK undertook the data collection and analysis process. KK drafted the initial manuscript, which was subsequently reviewed and commented on by RM. All authors participated in the review process and provided final approval for the manuscript.

Funding The authors declare that no funds, grants, or other support were received for the conduct of this study and preparation of this manuscript.

Data availability The PROMISE, RELINK, NASA and AEEEM datasets used in this study are publicly available at (https://zenodo.org/records/3693686), (https://www.cse.ust.hk/~scc/RELINK.htm), (https://github.com/klainfo/NASADefectDataset) and (https://bug.inf.usi.ch/download.php) respectively.

Declarations

Competing interests The authors declare no competing interests.

Ethical approval Not applicable.

Human participants and/or animals Not applicable.

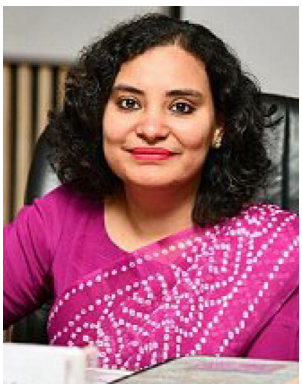
References

1. Ramulu, K., Murhtyr, B.V.R.: Importance of software quality models in software engineering. Int. J. Eng. Technol. Manag. Res. **5**, 200–218 (2020). https://doi.org/10.29121/ijetmr.v5.i3.2018.192
2. Miguel, J.P., Mauricio, D., Rodríguez, G.: A review of software quality models for the evaluation of software products. Int. J. Softw. Eng. Appl. **5**, 31–53 (2014). https://doi.org/10.5121/ijsea.2014.5603
3. Yu, Q., Jiang, S., Qian, J.: Which is more important for cross-project defect prediction: instance or feature? In: 2016 International Conference on Software Analysis, Testing and Evolution (SATE). IEEE, pp. 90–95 (2016)
4. Balogun, A.O., Basri, S., Mahamad, S., et al.: Impact of feature selection methods on the predictive performance of software defect prediction models: an extensive empirical study. Symmetry **12**, 1147 (2020). https://doi.org/10.3390/sym12071147
5. Kakkar, M., Jain, S.: Feature selection in software defect prediction: a comparative study. In: 2016 6th International Conference—Cloud System and Big Data Engineering (Confluence). IEEE, pp 658–663 (2016)
6. Abdel-Basset, M., El-Shahat, D., El-henawy, I., et al.: A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection. Expert Syst. Appl. **139**, 112824 (2020). https://doi.org/10.1016/j.eswa.2019.112824
7. Kassaymeh, S., Abdullah, S., Al-Betar, M.A., Alweshah, M.: Salp swarm optimizer for modeling the software fault prediction problem. J. King Saud Univ.-Comput. Inf. Sci. **34**, 3365–3378 (2022). https://doi.org/10.1016/j.jksuci.2021.01.015
8. Ezugwu, A.E., Shukla, A.K., Nath, R., et al.: Metaheuristics: a comprehensive overview and classification along with bibliometric analysis. Artif. Intell. Rev. **54**, 4237–4316 (2021). https://doi.org/10.1007/s10462-020-09952-0
9. Ferenc, R., Toth, Z., Ladanyi, G., et al.: A public unified bug dataset for java and its assessment regarding metrics and bug prediction. Softw. Qual. J. **28**, 1447–1506 (2020). https://doi.org/10.1007/s11219-020-09515-0
10. Jureczko, M., Madeyski, L.: Towards identifying software project clusters with regard to defect prediction. In: Proceedings of the 6th International Conference on Predictive Models in Software Engineering. ACM, pp. 1–10 (2010)
11. Wu, R., Zhang, H., Kim, S., Cheung, S.-C.: RELINK: recovering links between bugs and changes. In: Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering, pp. 15–25 (2011)
12. D'Ambros, M., Lanza, M., Robbes, R.: Evaluating defect prediction approaches: a benchmark and an extensive comparison. Empir. Softw. Eng. **17**, 531–577 (2012). https://doi.org/10.1007/s10664-011-9173-9
13. Mirjalili, S., Mirjalili, S.M., Lewis, A.: Grey wolf optimizer. Adv. Eng. Softw. **69**, 46–61 (2014). https://doi.org/10.1016/j.advengsoft.2013.12.007
14. El-Hasnony, I.M., Barakat, S.I., Mostafa, R.R.: Optimized ANFIS model using hybrid metaheuristic algorithms for Parkinson's disease prediction in IoT environment. IEEE Access **8**, 119252–119270 (2020). https://doi.org/10.1109/ACCESS.2020.3005614
15. Mirjalili, S., Lewis, A.: S-shaped versus V-shaped transfer functions for binary Particle Swarm Optimization. Swarm Evol. Comput. **9**, 1–14 (2013). https://doi.org/10.1016/j.swevo.2012.09.002
16. Goyal, S.: Genetic evolution-based feature selection for software defect prediction using SVMs. J. Circ. Syst. Comput. **31**, 2250161 (2022). https://doi.org/10.1142/S0218126622501614
17. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
18. Gezici, B., Tarhan, A.K.: Explainable AI for software defect prediction with gradient boosting classifier. In: 2022 7th International Conference on Computer Science and Engineering (UBMK), pp. 1–6 (2022)
19. Karim, S., Leslie Hendric Spits Warnars, H., Lumban Gaol, F., et al.: Software metrics for fault prediction using machine learning approaches a literature review with PROMISE repository dataset. In: 2017 IEEE International Conference on Cybernetics and Computational Intelligence (Cybernetics Com), pp. 19–23 (2017)
20. Dahim, M., Alqadhi, S., Mallick, J.: Enhancing landslide management with hyper-tuned machine learning and deep learning models: predicting susceptibility and analyzing sensitivity and uncertainty. Front. Ecol. Evol. **11**, 1108924 (2023). https://doi.org/10.3389/fevo.2023.1108924
21. Malik, N.U.R., Sheikh, U.U., Abu-Bakar, S.A.R., Channa, A.: Multi-view human action recognition using skeleton based-FineKNN with extraneous frame scrapping technique. Sensors **23**, 2745 (2023). https://doi.org/10.3390/s23052745
22. Malhotra, R., Khan, K.: A study on software defect prediction using feature extraction techniques. In: 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE, pp. 1139–1144 (2020)
23. Heidari, A.A., Mirjalili, S., Faris, H., et al.: Harris hawks optimization: Algorithm and applications. Future Gener. Comput. Syst. **97**, 849–872 (2019). https://doi.org/10.1016/j.future.2019.02.028
24. Khatri, Y., Singh, S.K.: An effective feature selection based cross-project defect prediction model for software quality improvement. Int. J. Syst. Assur. Eng. Manag. **14**, 154–172 (2023). https://doi.org/10.1007/s13198-022-01831-x

25. Mirjalili, S., Lewis, A.: The whale optimization algorithm. Adv. Eng. Softw. **95**, 51–67 (2016). https://doi.org/10.1016/j.advengsoft.2016.01.008
26. da Silva, L.S.A., Lúcio, Y.L.S., dos Coelho, L.S., et al.: A comprehensive review on Jaya optimization algorithm. Artif. Intell. Rev. **56**, 4329–4361 (2023). https://doi.org/10.1007/s10462-022-10234-0
27. Sharma, T., Sangwan, O.P.: Sine-Cosine algorithm for software fault prediction. In: Proceedings—2021 IEEE International Conference on Software Maintenance and Evolution, ICSME 2021. Institute of Electrical and Electronics Engineers Inc., pp. 701–706 (2021)
28. Riaz, S., Arshad, A., Jiao, L.: Rough noise-filtered easy ensemble for software fault prediction. IEEE Access **6**, 46886–46899 (2018). https://doi.org/10.1109/ACCESS.2018.2865383
29. Tumar, I., Hassouneh, Y., Turabieh, H., Thaher, T.: Enhanced binary moth flame optimization as a feature selection algorithm to predict software fault prediction. IEEE Access **8**, 8041–8055 (2020). https://doi.org/10.1109/ACCESS.2020.2964321
30. Brundha Elci, J., Nandagopalan, S.: SS-WDRN: sparrow search optimization-based weighted dual recurrent network for software fault prediction. Knowl. Inf. Syst. **66**, 1037–1064 (2024). https://doi.org/10.1007/s10115-023-01932-4
31. Sekaran, K., Lawrence, S.P.A.: Mutation boosted salp swarm optimizer meets rough set theory: a novel approach to software defect detection. Trans. Emerg. Telecommun. Technol. (2024). https://doi.org/10.1002/ett.4953
32. Das, H., Prajapati, S., Gourisaria, M.K., et al.: Feature selection using golden jackal optimization for software fault prediction. Mathematics **11**, 2438 (2023). https://doi.org/10.3390/math11112438
33. Tang, Y., Dai, Q., Yang, M., et al.: Software defect prediction ensemble learning algorithm based on adaptive variable sparrow search algorithm. Int. J. Mach. Learn. Cybern. (2023). https://doi.org/10.1007/s13042-022-01740-2
34. Mafarja, M., Thaher, T., Al-Betar, M.A., et al.: Classification framework for faulty-software using enhanced exploratory whale optimizer-based feature selection scheme and random forest ensemble learning. Appl. Intell. **53**, 18715–18757 (2023). https://doi.org/10.1007/s10489-022-04427-x
35. Zada, I., Alshammari, A., Mazhar, A.A., et al.: Enhancing IOT based software defect prediction in analytical data management using war strategy optimization and Kernel ELM. Wirel. Netw. (2023). https://doi.org/10.1007/s11276-023-03591-3
36. Alweshah, M., Kassaymeh, S., Alkhalaileh, S., et al.: An efficient hybrid mine blast algorithm for tackling software fault prediction problem. Neural. Process. Lett. **55**, 10925–10950 (2023). https://doi.org/10.1007/s11063-023-11357-3
37. Gupta, A., Sharma, M., Srivastava, A.: A novel dimensionality reduction-based software bug prediction using a bat-inspired algorithm. In: Proceedings of the 13th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2023. Institute of Electrical and Electronics Engineers Inc., pp. 278–285 (2023)
38. Sahu, S.P., Reddy, B.R., Mukherjee, D., et al.: A hybrid approach to software fault prediction using genetic programming and ensemble learning methods. Int. J. Syst. Assur. Eng. Manag. **13**, 1746–1760 (2022). https://doi.org/10.1007/s13198-021-01532-x
39. Balaram, A., Vasundra, S.: Prediction of software fault-prone classes using ensemble random forest with adaptive synthetic sampling algorithm. Autom. Softw. Eng. **29**, 6 (2022). https://doi.org/10.1007/s10515-021-00311-z
40. Thirumoorthy, K., Jerold John Britto, J.: A feature selection model for software defect prediction using binary Rao optimization algorithm. Appl. Soft Comput. **131**, 109737 (2022). https://doi.org/10.1016/j.asoc.2022.109737
41. Thirumoorthy, K., Britto, J.J.J.: A clustering approach for software defect prediction using hybrid social mimic optimization algorithm. Computing **104**, 2605–2633 (2022). https://doi.org/10.1007/s00607-022-01100-6
42. Zhu, K., Ying, S., Zhang, N., Zhu, D.: Software defect prediction based on enhanced metaheuristic feature selection optimization and a hybrid deep neural network. J. Syst. Softw. **180**, 111026 (2021). https://doi.org/10.1016/j.jss.2021.111026
43. Hassouneh, Y., Turabieh, H., Thaher, T., et al.: Boosted whale optimization algorithm with natural selection operators for software fault prediction. IEEE Access **9**, 14239–14258 (2021). https://doi.org/10.1109/ACCESS.2021.3052149
44. Thaher, T., Arman, N.: Efficient multi-swarm binary harris hawks optimization as a feature selection approach for software fault prediction. In: 2020 11th International Conference on Information and Communication Systems, ICICS 2020. Institute of Electrical and Electronics Engineers Inc., pp 249–254 (2020)
45. Kalavani, N., Beena, R.: Boosted relief feature subset selection and heterogeneous cross project defect prediction using firefly particle swarm optimization. Int. J. Recent Technol. Eng. **8**, 2605–2613 (2020). https://doi.org/10.35940/ijrte.E6333.018520
46. Anbu, M., Anandha Mala, G.S.: Feature selection using firefly algorithm in software defect prediction. Clust. Comput. **22**, 10925–10934 (2019). https://doi.org/10.1007/s10586-017-1235-3
47. Turabieh, H., Mafarja, M., Li, X.: Iterated feature selection algorithms with layered recurrent neural network for software fault prediction. Expert Syst. Appl. **122**, 27–42 (2019). https://doi.org/10.1016/j.eswa.2018.12.033
48. Khuat, T.T., Le, M.H.: Binary teaching–learning-based optimization algorithm with a new update mechanism for sample subset optimization in software defect prediction. Soft Comput. **23**, 9919–9935 (2019). https://doi.org/10.1007/s00500-018-3546-6

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Ruchika Malhotra is the Head of the Department and a Professor in the Department of Software Engineering at Delhi Technological University (DTU), Delhi, India. She has previously served as the Associate Dean of Industrial Research and Development at DTU. She was awarded the prestigious Raman Fellowship for post-doctoral research at Indiana University-Purdue University Indianapolis, USA. She earned her master's and doctorate degrees in software engineering from the University School of Information Technology at Guru Gobind Singh Indraprastha University, Delhi, India. In 2013, she received the IBM Faculty Award. Her contributions to the field have earned her recognition as one of the world's top 2% scientists, according to a Stanford University report in 2020, specifically for her work in "Artificial Intelligence & Image Processing" (PLOS

Biology, 2020). She has also received the Commendable Research Award from Delhi Technological University for the years 2018-2022. Her h-index is 37 as reported by Google Scholar. She is the author of the book "Empirical Research in Software Engineering," published by CRC Press, and co-author of "Object Oriented Software Engineering," published by PHI Learning. She has published over 200 research papers in international journals and conferences. Her research interests include software testing, software quality improvement, statistical and adaptive prediction models, software metrics, and the definition and validation of software metrics.

Kishwar Khan received her B.Tech. degree in Computer Science and Engineering from Dr. A. P. J. Abdul Kalam Technical University, Lucknow, India, and her M.Tech. degree in Software Engineering from Delhi Technological University, Delhi, India. She is currently pursuing a Ph.D. in Software Engineering at Delhi Technological University. Her doctoral research focuses on software quality improvement and the application of machine learning and deep learning in various aspects of software quality. Her current research interests include Artificial Intelligence, software quality, predictive modelling, and data analytics.

Check for updates

A Single Bitline Highly Stable, Low Power With High Speed Half-Select Disturb Free 11T SRAM Cell

LOKESH SONI, Delhi Technological University, Delhi, India
NEETA PANDEY, Delhi Technological University, Delhi, India

A half-select disturb-free 11T (HF11T) static random access memory (SRAM) cell with low power, better stability and high speed is presented in this paper. The proposed SRAM cell works well with bit-interleaving design, which enhances soft-error immunity. A comparison of the proposed HF11T cell with other cutting-edge designs such as single-ended HS free 11T (SEHF11T), a shared-pass-gate 11T (SPG11T), data-dependent stack PMOS switching 10T (DSPS10T), a single-ended half-selected robust 12T (HSR12T), and 11T SRAM cells has been made. It exhibits 4.85×/9.19× less read delay (T_{RA}) and write delay (T_{WA}), respectively as compared to other considered SRAM cells. It achieves 1.07×/1.02× better read and write stability, respectively than the considered SRAM cells. It shows maximum reduction of 1.68×/4.58×/94.72×/9×/145× leakage power, read power, write power consumption, read power delay product (PDP) and write PDP respectively, than the considered SRAM cells. In addition, the proposed HF11T cell achieves 10.14× higher I_{on}/I_{off} ratio than the other compared cells. These improvements come with a trade-off, resulting in 1.13× more T_{RA} compared to SPG11T. The simulation is performed with Cadence Virtuoso 45nm CMOS technology at supply voltage (V_{DD}) of 0.6 V.

CCS Concepts: • **Hardware** → **Integrated circuits**; **Semiconductor memory**; **Static memory**.

Additional Key Words and Phrases: Single-bitline SRAM, half-select disturb-free, read and write delay, read static noise margin (RSNM), hold static noise margin (HSNM) and write ability, power consumption, I_{on}/I_{off} ratio

1 Introduction

Power dissipation has emerged as a crucial design constraint due to the rise in demand for low power devices pertaining to wireless sensor networks, biomedical devices, and other battery-operated portable devices. The power dissipation in Systems-on-Chip (SoCs) is primarily influenced by Static Random Access Memory (SRAM) since it occupies a significant portion of the chip. Moreover, the proportion of SRAM is expected to increase even more in the future [1]. Therefore, designing of low power SRAM cell is important. To extend the battery life in SRAM cells and reduce power dissipation, the most efficient approach is to decrease the supply voltage (V_{DD}). This downscaling of V_{DD} proves to be the most effective method. The reason behind this is that, standby power decreases exponentially and dynamic power decreases quadratically as the V_{DD} is reduced [2]. However, process variation significantly limits the performance of the SRAM cell at lower supply voltages. Therefore, it is challenging to maintain the device strength ratio in conventional 6T SRAM cell leading to read/write stability issue at low V_{DD} [3]. Various SRAM cells have been proposed to address the read stability issue [4-9]. These cells provide significant improvements in read static noise margin (RSNM) by employing the separate read/write path, but they have poor write margin and more power consumption [4-7,9] due to the presence of dual bitline. Additionally, a number of write-assist methods have been discussed in the literature to boost the SRAM cell's

Just Accepted

Authors' Contact Information: Lokesh Soni, Delhi Technological University, Delhi, Delhi, India; e-mail: lokeshsoni_2k21phdec502@dtu.ac.in; Neeta Pandey, Delhi Technological University, Delhi, India; e-mail: n66pandey@rediffmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 1557-7309/2024/6-ART
https://doi.org/10.1145/3653675

write ability [10-12]. Wordline (WL) boosting [10], [11], and negative bitline (NBL) [12] are popularly used write assist methods for enhancing the write ability. On the other hand, soft errors become more prominent with the technology scaling, that affect the stability of a cell [13]. To overcome this a bit-interleaving architecture is commonly used. However, this method is applicable to the cells that are fully half-select (HS) free in nature. In order to address HS issue, different approaches have been suggested, by employing the two distinct wordlines (WLs) [14-20]. Though, the use of serially coupled transistors in the write path, the designs from [14-20] have poor write ability. As a result, various write-assist strategies, like WL boosting or negative BL, are necessary for effective write operation, but these techniques having a large area overhead and more dynamic power consumption. The cells in [14-17] consumes more power due to the presence of two bitlines (BLs) for read and write operations. Although, the single BL cells present in [18-20] shows low dynamic power consumption but they have poor write and read ability and more leakage power consumption. The read ability and leakage power are more since two access transistors are connected to BL [18] that may lead to more BL capacitance, whereas one of the two series connected transistor in [19-20] are getting off that will cause charge sharing. Other single BL SRAM cells using shared-pass-gate [21] and power cut-off with floating avoidance [22] to remove HS issue are presented in literature. However, [21,22] shows the more write delay and power consumption due to the presence of more number of transistors in their write path and large BL capacitance, respectively.

In this paper, a single bitline half-select disturb-free 11T (HF11T) SRAM cell is proposed. This cell shows high read and write stability, less write delay, lower power consumption and better I_{on}/I_{off} ratio. The key characteristics of the suggested cell are outlined below:

(1) The RSNM is improved by using a read decoupled structure.
(2) The write delay (T_{WA}) and stability is improved by employing a power cut-off transistor with virtual ground (VGND) along with dedicated write path for '1' and '0'.
(3) By utilising a single bitline (BL) structure, it's dynamic power consumption is reduced.
(4) Leakage current and power are reduced with the help of stacked transistor, leading to better I_{on}/I_{off} ratio.
(5) It is free from Half-select-issue. As a result, bit-interleaving architecture can be used to improve resilience against soft errors.

This paper is organized as follows: Section II, describes the structure and operation of the proposed HF11T SRAM cell. In Section III, simulation results and analysis for the proposed and considered single BL SRAM cells are presented. Finally, conclusion are drawn in Section IV.

2 Proposed HF11T SRAM cell Design

Fig. 1(a), depicts the proposed HF11T SRAM cell. It is made up with a cross-coupled structure of a right inverter (P2, N2), and a left inverter with stacked transistors (P3, P1, N1) with VGND. The P3 is used as a power cut-off transistor at the time of write operation. The P5 and N4 are used for controlling the write access transistors N3 and P4 through the row and column control signals WL, CWL and WLA, respectively. The read access path consist of N5 and N6 transistors controlled by read wordline (RWL) signal. The VGND signal is used to control the write operation, whereas Q and QB are the storage nodes of a SRAM cell. The operating condition for hold, read, and write modes are enlisted in Table 1 and timing diagram is depicted in Fig. 1(b).

2.1 Hold Operation

Fig. 2(a) shows the schematic for the hold mode. The signals WL, WLA, RWL and VGND are maintained at '0', whereas CWL and BL are at '1', respectively. As a result, N3, P4, N4, and N5 are turned off, which makes the storage node Q isolated from the BL. Therefore, the cell successfully retains the data.

Just Accepted

Table 1. HF11T SRAM Cell Control Signals

Signal	Hold	Read	Write '1'	Write '0'
WL	0	0	0	1
WLA	0	0	1	0
CWL	1	1	0	1
RWL	0	1	0	0
VGND	0	0	1	0
BL	1	Precharged	1	0

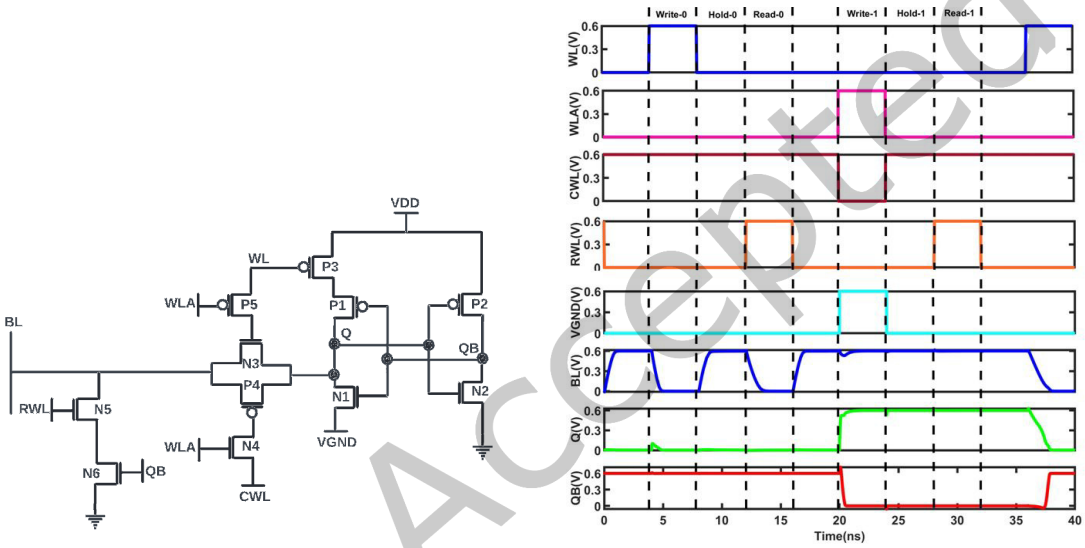


Fig. 1. Proposed HF11T SRAM cell (a) Schematic and (b) operation timing diagram.

Just Accepted

2.2 Read Operation

The read operation is elucidated in Fig. 2(b). BL is precharged to V_{DD} prior to starting a read operation. At the time of read operation WL, WLA and VGND are kept at '0'; CWL and RWL are kept at '1'. This makes N3, N4 and N5 are turned off and on, respectively. If the data stored in cell is '0' then QB is at '1' which activates N6. It allows discharging of BL through N5-N6 which is interpreted as '0' stored in the cell. On the other hand no discharging of BL indicates that stored value in cell is '1'.

2.3 Write Operation

The HF11T SRAM cell write '1' operation is shown in Fig. 2(c). Initially '0' and '1' are stored at Q and QB respectively. For write '1' operation WL, CWL and RWL are kept at '0', while WLA and VGND are connected to '1', and write driver makes BL to '1'. As P3 is on it connects the left inverter to the power supply (V_{DD}). The write access transistor P4 (PMOS) and N3 (NMOS) are turned on and off, respectively. As PMOS passes strong '1', the

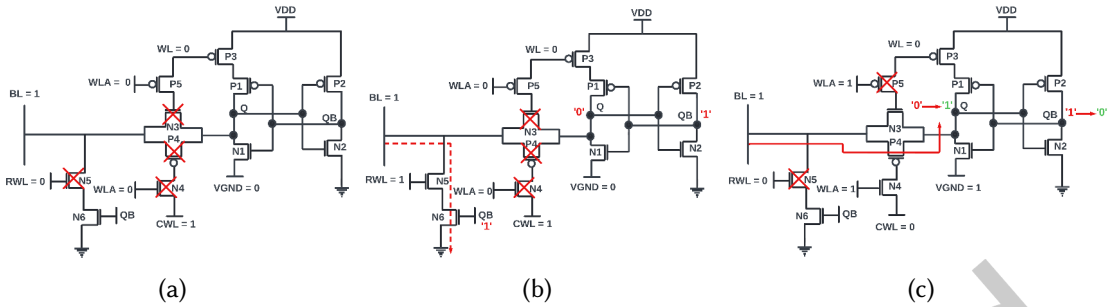


Fig. 2. (a) Hold, (b) Read, and (c) Write '1' operations of proposed HF11T cell

data at storage node Q get easily flipped from '0' to '1'. Also, P3 and VGND support write '1' operation, therefore the write ability of the proposed cell is improved.

Similarly, assume that '1' is stored at Q, so while performing a write '0' operation, BL is set to '0' by the write driver. The control signals WLA, RWL and VGND are disabled, whereas WL and CWL are enabled. The left inverter is disconnected from the V_{DD} as P3 is turned off. N3 and P4 are switched on and off respectively, which forming a write path. Since NMOS act as strong for '0', Q is easily pulled low by BL via N3. As it reaches the threshold voltage of P2, QB is pulled up, which represent a successful write '0' operation. Furthermore, the write '0' operation is supported by P3 and VGND. Therefore, the above discussion shows that write operation of the proposed cell is improved with the help of power cut-off transistor, VGND, and strong write access transistor depending on write '1' & '0' operation.

2.4 Write-Half-Select Free Operation

A disturbance created by any unselected cells in selected rows or columns during a writing process is known as a half-select (HS) disturb [22]. In order to address the multi-bit errors HS free SRAM cell is required to implement BI architecture. The proposed HF11T completely removes HS issue as shown in Fig. 3. There are four bit-cells, each of which represents one of four possible conditions namely Selected cell, Row half-selected cell (RHS), Column half-selected cell (CHS) and Unselected cell during write operation. The unselected and selected cells are working in normal hold and write modes as explained in Section II-A and II-C, respectively. As P5, N3 and P4 are deactivated for the CHS cell, the storage nodes are separated from BL, so that the latch consistently holds the data just as it would in a hold operation. For RHS cell, WL, CWL and BL are at '0', whereas WLA and VGND are at '1'. Assuming that, if the cell stores '1', then Q will remain at '1'. Since the strong pull-up path prevents it from being discharged. As a result, QB keeps at '0', allowing for secure latching of the stored data. In case of '0' stored in a cell, due to an existing discharging path from Q to ground via P4, Q state is unaffected and remains at '0'. As a result, the data stored in RHS cell is unaffected like hold state. In conclusion, the proposed SRAM cell is free from write-half-select disturb issue.

3 Simulation Results And Analysis

The performance of the proposed HF11T SRAM cell in terms of read, hold and write stability, read and write delay, read and write power, leakage power, and I_{on}/I_{off} ratio is evaluated using a 45 nm CMOS technology in Cadence Virtuoso. The V_{DDmin} of all the compared single BL SRAM cells namely SEHF11T [18], SPG11T [21], DSPS10T [19], HSR12T [20], 11T [22] and proposed HF11T SRAM cells is examined and the maximum of V_{DDmin} (0.6V) is chosen for further exploration. For fair comparison the minimum width of each transistor i.e. 120 nm, is taken

Just Accepted

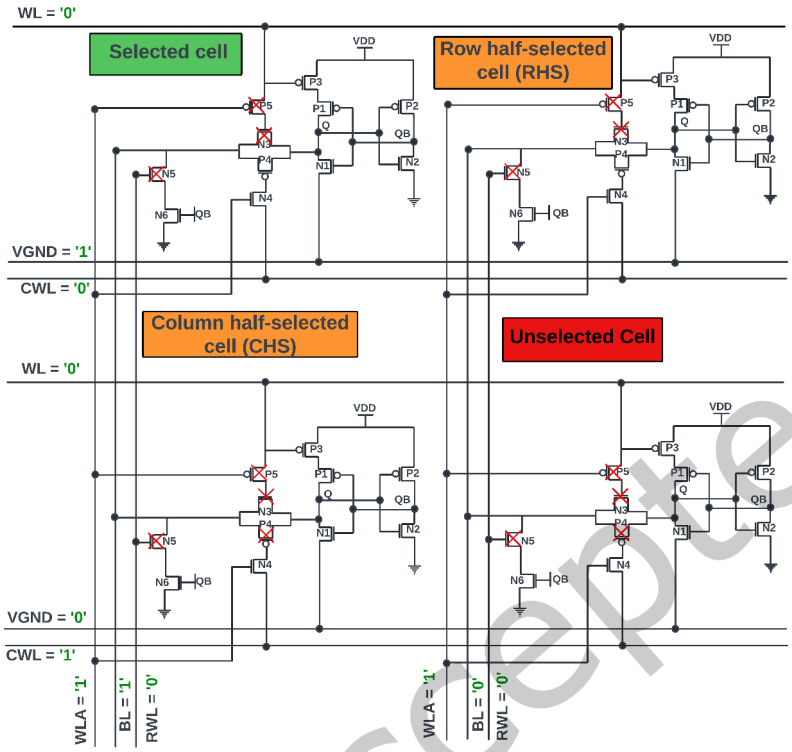


Fig. 3. HF11T SRAM cell resolves the issue of write-half-select disturb.

into account when simulating and comparing all the considered cells. The performance metrics of the proposed and compared bit cells are analyzed by conducting 5000 Monte Carlo (MC) simulations, used to determine the effect of process variations. The outcomes are discussed separately in the following subsections.

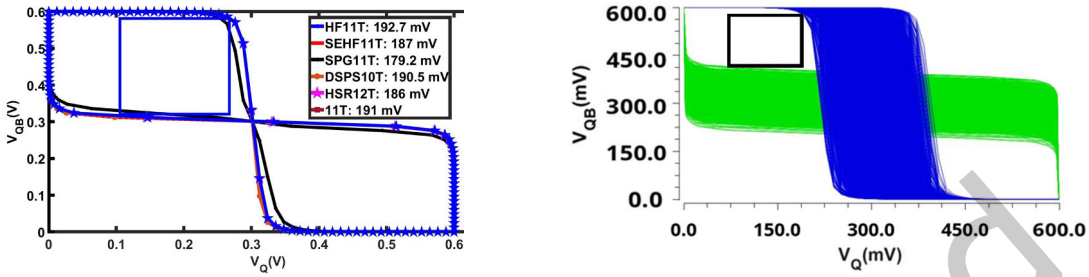
3.1 Read/Hold Static Noise Margin (RSNM/HSNM)

The RSNM and HSNM are used to determine the read and hold stability of a SRAM cell's, respectively. During read and hold operations, RSNM and HSNM are defined as the maximum dc noise voltage at each storage node that can be tolerated without affecting the status of stored data. They are calculated using the largest square that can fit inside the butterfly curve [23,24]. The read decoupled structure is employed in the proposed and compared SRAM cells, which having independent read path, as a result, their RSNM and HSNM are equivalent.

The RSNM/HSNM curve for all the compared SRAM cells are shown in Fig. 4(a). It can be seen from Fig. that the RSNM/HSNM of the proposed HF11T is 192.7 mV which is about 1.07× more than the other considered cell's. The proposed cell features stacked PMOS in the pull-up path, combined with the left-side structure, resulting in increased resistance toward the storage node. This configuration helps control noise, leading to a higher RSNM. The SEHF11T, DSPS10T, HSR12T, and 11T shows comparable RSNM to the proposed cell. The SEHF11T, and DSPS10T are having same kind of pull up structure and HSR12T, and 11T shows the similar pull up network with feedback cutting with stacked PMOS. This increases the equivalent resistance, contributing to the control of noise at the storage nodes during a read/hold operation. However, SPG11T doesn't have the stacked PMOS in

Just Accepted

the pull up path leading to smaller RSNM than the other cells. The MC simulation for 5000 samples of HF11T is illustrated in Fig. 4(b). It indicates a worst-case RSNM/HSNM of about 110 mV.



(a) (b)

Fig. 4. RSNM (a) Butterfly curve for compared SRAM cells, (b) of HF11T using MC 5000 runs at V_{DD} = 0.6 V.

3.2 Write ability

The N-Curve [25,26], which is depicted in Fig. 5(a), is used to evaluate the write capability of the SRAM cell's. The write stability of bitcells can be measured using two different metrics: write-trip voltage (WTV) and write-trip current (WTI). The WTV, which is defined by the difference between points C and B (Fig. 5(a)), is the smallest voltage drop necessary to modify the internal nodes of the bitcell. The negative current peak between points C and B in Fig. 5(a), used to determine the WTI, which represents the minimum current required to write the bitcell. Fig. 5(b) depicts the N-Curve corresponding to each SRAM cell. It can be seen that the WTV of the proposed HF11T cell is 1.02× better than the considered cells. The power cut-off transistor with VGND and a dedicated write path for '0' and '1' improves the write capability of HF11T. Fig. 5(c) demonstrates that the worst-case WTV for the proposed cell is 370 mV, as determined by MC simulation with 5000 samples at V_{DD} of 0.6 V.

3.3 Read Delay

Read delay/access time (T_{RA}) is used to measure the time required to complete a read operation. For single ended bit cells, it is determined as the time needed to discharge precharged BL by 50 mV when WL is activated [25,27,31]. The T_{RA} of various SRAM cells at varying supply voltages are displayed in Fig. 6(a). As shown HF11T has 1.48×/4.85×/1.92×/1.64× lower read delay than SEHF11T/ DSPS10T/ HSR12T/ 11T SRAM cells, respectively at V_{DD} of 0.6 V, due to reduced bitline capacitance. However, it has slightly more read delay than SPG11T, because of bit large BL capacitance. The DSPS10T and HSR12T shows maximum delay due to voltage division occurs at the intermediate node between two series connected access transistors and decoupled read path. In the 11T cell, the node QX gets charged before the activation of the WL signal. Upon WL activation, QX is rapidly discharged, causing a swift drop on RBL. Consequently, a smaller delay is observed compared to previous cells. The SEHF11T shows more read delay due to large capacitance offered by two parallelly connected access transistors with single BL. Fig. 6(b) illustrates the T_{RA} at V_{DD} of 0.6 V with MC 5000 runs. The proposed cell exhibits a lower mean (μ) and minimum standard deviation (σ) value of up to 4.85× and 10.53×, respectively than other considered cells.

3.4 Write Delay

Write delay/access time (T_{WA}) shows the speed at which it executes the write operation. The amount of time an internal node needs to increase up to 90% of the supply voltage while the wordline is enabled is referred to as the write '1' access time (T_{WA}'1'). Similarly, write '0' access time (T_{WA}'0') is defined as the time when the storage

Just Accepted

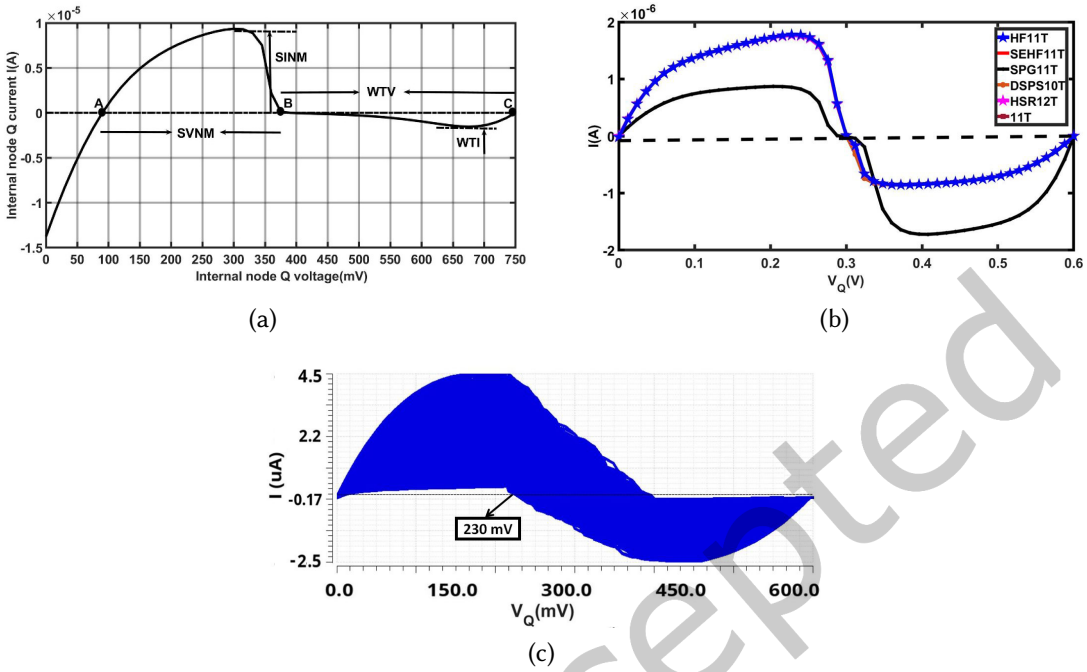


Fig. 5. Representation of N-curve for (a) conventional 6T SRAM cell, (b) compared SRAM cells, (c) MC simulation of HF11T with 5000 runs at V_{DD} = 0.6V.

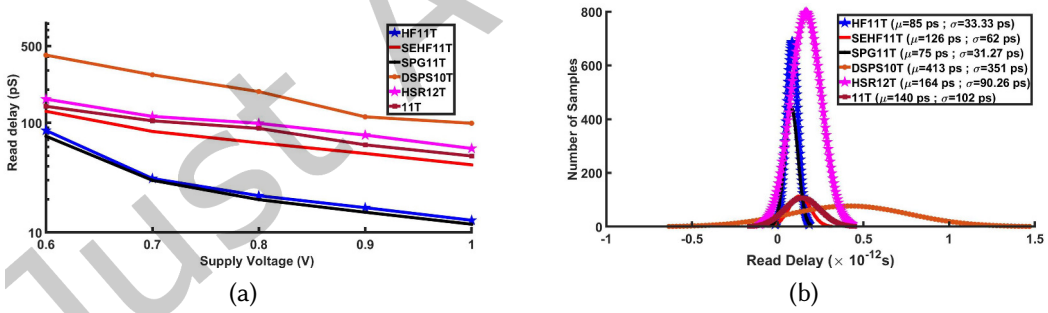


Fig. 6. Read delay (a) versus varying supply voltages, (b) with MC simulation of 5000 samples.

node voltage drops upto 10% of the supply voltage during wordline activation [25,28]. Fig. 7(a) shows the T_{WA} of considered SRAM cells at various supply voltages. Because of the power cut-off transistor with VGND and dedicated write path for '1' and '0', HF11T has 8.81×/1.12×/9.19×/8.18×/1.35× lower write delay than SEHF11T/ SPG11T/ DSPS10T/ HSR12T/ 11T SRAM cells, respectively at V_{DD} of 0.6 V. MC simulation with 5000 samples at V_{DD} of 0.6 V is shown in Fig. 7(b). The proposed cell indicates a lower mean (μ) and minimum standard deviation (σ) of up to 9.19× and 16.36×, respectively than other considered cells.

Just Accepted

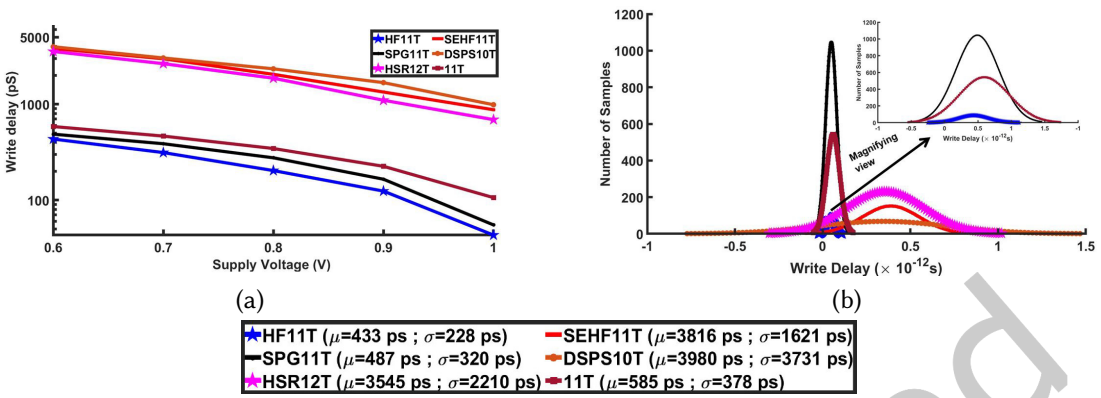


Fig. 7. Write delay (a) versus varying supply voltages, (b) with MC simulation of 5000 samples.

3.5 Read/Write Power Consumption

Read power is calculated as the average power during read operations. In similar way, write power is determine as the average power during write operations including write '1' and '0' [21]. Variation in read and write power with supply voltage is shown in Fig. 8(a) & (b), respectively. From Fig., HF11T has 1.59×/2.75×/1.77×/4.58×/3.84× and 16.35×/94.72×/13.55×/3.48×/60.81× lesser read and write power consumption than SEHF11T/ SPG11T/ DSPS10T/ HSR12T/ 11T SRAM cells, respectively at V_{DD} of 0.6 V. The proposed cell shows lower read and write power consumption due to less BL capacitance. The SEHF11T consumes more read and write power this occurs because of two transistors are linked to the bitline resulting in an overall increase in the bitline capacitance. The HSR12T, 11T, and SPG11T consumes maximum read power owing to the substantial read current flow and due to the introduction of P2, which gives rise to a larger BL capacitance, respectively; whereas DSPS10T shows less read power than these cells due to minimal flow of read current in the read path. Also, the SPG11T and 11T exhibits the highest write power among the mentioned cells, primarily attributed to the presence of the write assist transistor and scheme; whereas DSPS10T and HSR12T shows less write power consumption due to larger delay as delay and power are inversely related to each other. MC simulations of read and write power with 5000 samples at V_{DD} of 0.6 V is shown in Fig. 9(a) & (b), respectively. In comparison to considered cells, the proposed cell indicates the lower mean (μ) and minimum standard deviations (σ) for read and write power of up to 4.58× and 8.10× and 94.72× and 97.48× respectively.

Fig. 10(a) & (b) illustrates the read and write power-delay product (PDP) at various supply voltages, respectively. HSR12T has largest read PDP due to its more read power, whereas SEHF11T has more write PDP because of large write delay. However, HF11T has minimum read power and least write delay/power as a result, it shows lower read and write PDP as compared to other considered cells.

3.6 Leakage Power Consumption

Leakage power is another crucial parameter for SRAM design. Because majority of the SRAM cells in an SRAM array remain in standby mode when not activated [21,32]. Leakage power consumption of various SRAM cells with supply voltage variation is shown in Fig. 11(a). The outcomes demonstrate that HF11T has 1.63×/1.09×/1.65×/1.68×/1.66× lower power consumption than SEHF11T/ SPG11T/ DSPS10T/ HSR12T/ 11T SRAM cells, respectively, due to the presence of stacking transistors a smaller number of leakage path and negligible current flowing through their off transistors; whereas SPG11T has the lowest leakage power consumption among

Just Accepted

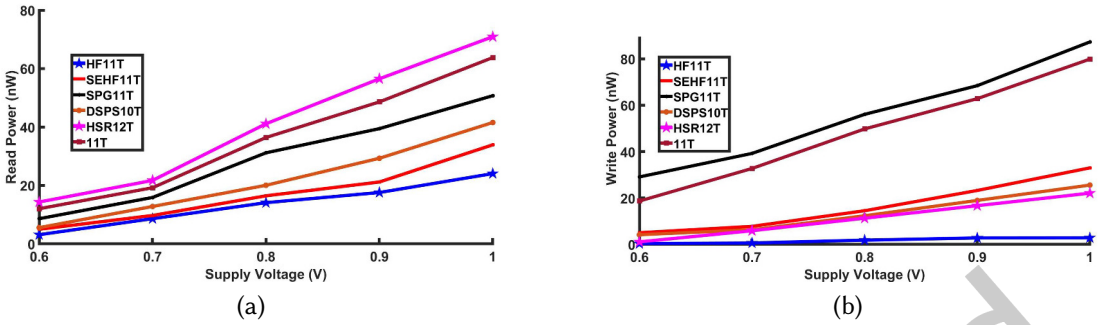


Fig. 8. Supply voltage variation on (a) Read power, and (b) Write power.

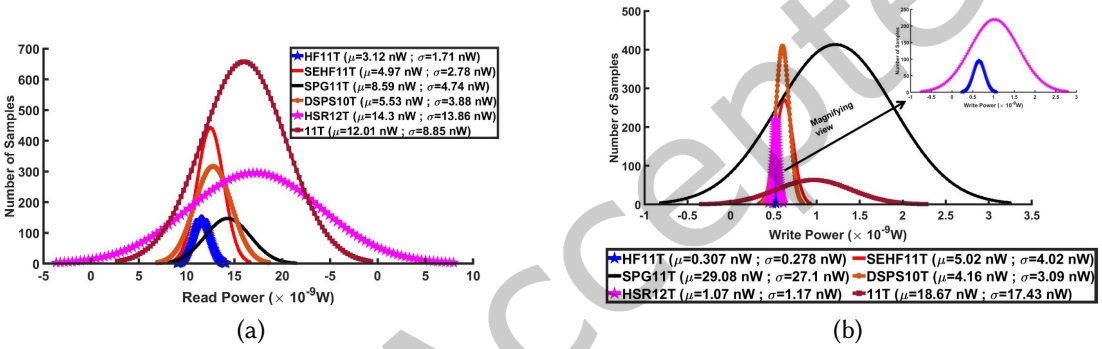


Fig. 9. MC simulations of 5000 samples for (a) Read power, and (b) Write power.

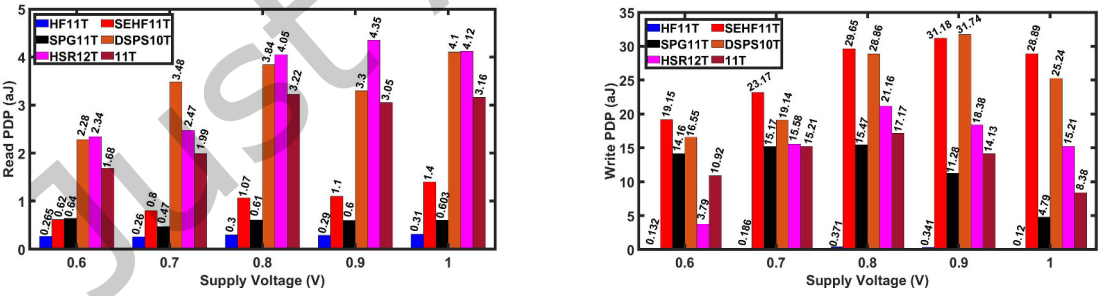


Fig. 10. Supply voltage variation on (a) Read PDP, and (b) Write PDP.

the compared cells due to low leakage through their off transistors. However, SEHF11T, DSPS10T, HSR12T, and 11T shows slightly more leakage power due to more leakage current flowing through their off transistors and more leakage path is present. Fig. 11(b) illustrate the MC simulation of leakage power consumption with 5000

Just Accepted

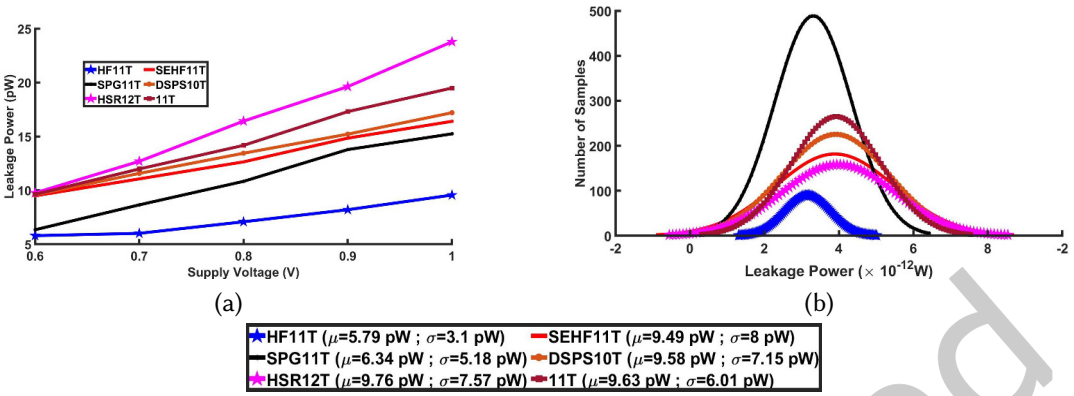


Fig. 11. Leakage power (a) versus supply voltage, (b) with MC simulation of 5000 samples.

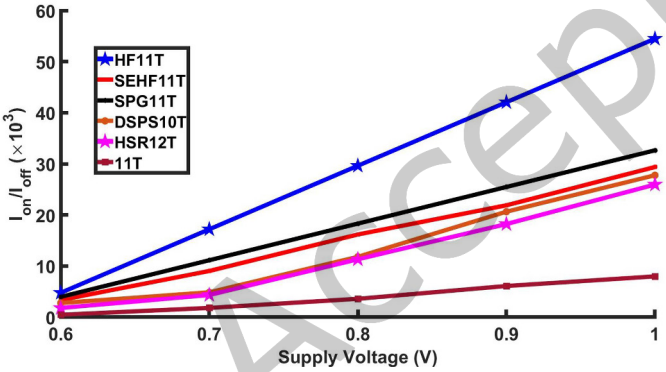


Fig. 12. I_{on}/I_{off} ratio variation with supply voltage.

samples at V_{DD} of 0.6 V. Compared to considered cells, the proposed cell indicates a lower mean (μ) and minimum standard deviation (σ) in leakage power consumption of up to 1.68× and 2.58×, respectively.

3.7 I_{on}/I_{off} ratio

The relationship between read current (I_{on}) and leakage current (I_{off}) is referred to as the I_{on}/I_{off} ratio. The designer can make a large SRAM array if the I_{on}/I_{off} ratio is higher. As a result, a large current ratio ensures that a designer can add more bit cells per column [25,29]. The I_{on}/I_{off} ratio with supply voltage variation is indicated in Fig. 12. The result shows that, HF11T has 1.42×/1.20×/1.72×/2.65×/10.14× more I_{on}/I_{off} ratio than SEHF11T/ SPG11T/ DSPS10T/ HSR12T/ 11T SRAM cells, respectively at V_{DD} of 0.6 V. The I_{on}/I_{off} ratio of the proposed cell is larger due to more I_{on} and less I_{off}, which is the direct consequence of stacking effect. The SEHF11T, DSPS10T, HSR12T and 11T shows least I_{on}/I_{off} ratio as two access transistors are connected in series, with one being off. This configuration introduces a potential for charge sharing between the two, leading to increased leakage current in the read path, whereas SPG11T shows slightly less I_{on}/I_{off} ratio due to the introduction of P2 in their read path causing the leakage.

Just Accepted

Table 2. Comparing several SRAM cells in 45-nm CMOS Technology.

Parameters	SEHF11T [18]	SPG11T [21]	DSPS10T [19]	HSR12T [20]	11T [22]	Proposed HF11T
V_{DD}	0.6	0.6	0.6	0.6	0.6	0.6
Read delay (ps)	126	**75**	413	164	140	85
Write delay (ps)	3816	487	3980	3545	585	**433**
RSNM (mv)	187	179.2	190.5	186	191	**192.7**
HSNM (mv)	187	179.2	190.5	186	191	**192.7**
WTV (mv)	299.66	307.16	299.66	299.18	299.19	**299.07**
WTI (µA)	-0.86	-1.72	-0.86	-0.85	-0.85	-0.86
Leakage Power (pw)	9.49	6.34	9.58	9.76	9.63	**5.79**
Read Power (nw)	4.97	8.59	5.53	14.3	12.01	**3.12**
Write Power (nw)	5.02	29.08	4.16	1.07	18.67	**0.30**
Read PDP (aJ)	0.62	0.64	2.28	2.34	1.68	**0.26**
Write PDP (aJ)	19.15	14.16	16.55	3.79	10.92	**0.13**
I_{on}/I_{off} **ratio** ($\times 10^3$)	3.35	3.97	2.77	1.8	0.47	**4.77**
Cell area (µm^2)	4.35	4.21	**3.55**	5.05	4.82	3.92
Relative FOM	0.0015	0.0018	0.0006	0.0016	0.00092	**1**

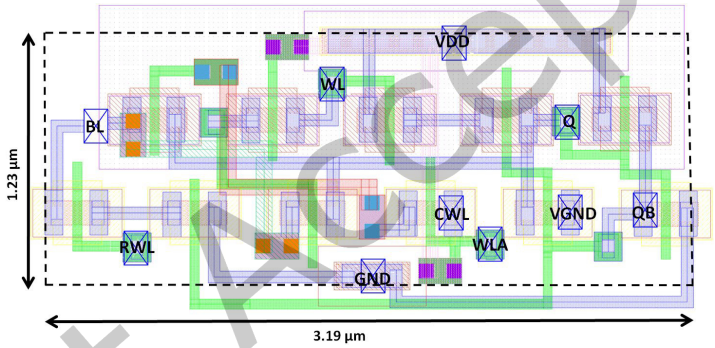


Fig. 13. Layout of HF11T.

Just Accepted

3.8 Cell Area

High density cache memory is created by grouping the bit cells, which requires layout area observation for each cell. Additionally, as cache memory occupy a significant amount of chip area, every memory designer aim to reduce the cell area while maintaining performance. The layouts were created using 45nm technology rules. The layouts of proposed and considered cells are drawn by keeping same widths and lengths of transistors. The area of all considered cells are placed in Table 2. The HF11T layout is shown in Fig. 13. The HF11T occupies 1.1×/1.07×/1.28×/1.22× lesser area than SEHF11T/ SPG11T/ HSR12T/ 11T SRAM cells, respectively. As SEHF11T and HSR12T, 11T has PMOS transistors positioned vertically on both sides and in feedback cutting path, respectively leading to more area requirement. The SPG11T has vertically connected PMOS on one side and contains more complex connections between the transistors leading to more area consumption. However, HF11T shows 1.1× more area consumption than DSPS10T. As HF11T has slightly more number of transistors than DSPS10T.

3.9 Figure of Merit (FOM)

In order to evaluate the proposed cell relative performance effectiveness with other compared SRAM cells under consideration, the SRAM figure of merit (FOM) presented in [30] has been used. The read delay (T_{RA}) and write delay (T_{WA}) are taken into account by this FOM, which is given in Eq. (1), as the speed metrics. For the hold, read, and write operations, respectively, the HSNM, RSNM, and WTV are taken into account as the cell's stability metrics.

$$FOM = \frac{RSNM \times HSNM}{WTV \times (T_{RA} \times T_{WA}) \times (Readpower \times Writepower \times Leakagepower) \times Cellarea} \quad (1)$$

The HF11T indicates $6.5\times10^2\times/5.4\times10^2\times/1.6\times10^3\times/5.8\times10^2\times/1\times10^3\times$ higher FOM than SEHF11T/ SPG11T/ DSPS10T/ HSR12T/ 11T SRAM cells, respectively. The difference in the FOM value of the proposed and SEHF11T [18], SPG11T [21], HSR12T [20] cells is nearly about 100 times. The large difference in the FOM values of the proposed and other compared cells [19,22] may be attributed to the fact that write delay and power of the proposed cell is approximately 100 times less than the compared cells respectively. A larger value of this FOM signifies improved cell performance.

To summarise the results, it can be seen from Table 2 that the HF11T performs better than any other SRAM cells compared in this paper.

4 Conclusion

A novel HF11T SRAM cell is proposed in this paper. This cell has power cut-off transistor with VGND along with dedicated write path for '1' and '0'. The proposed cell solves read disturb and write-half-select disturb problems. The HF11T results in low leakage power, low read and write power consumption, better read and write delays, better I_{on}/I_{off} ratio, and improved RSNM, HSNM and write ability. It achieves up to 1.68×/4.58×/94.72× reduction in leakage power, read power and write power consumption, respectively than other compared SRAM cells, as it contains stacking transistor, single BL for read/write and low BL capacitance. The write delay and stability of the proposed cell is up to 9.19×/1.02× better than other considered cells, as a result of power cut-off transistor with VGND and dedicated write path for '1' and '0'. Although, read delay is least among the compared SRAM cells but it shows slightly more delay than SPG11T. The I_{on}/I_{off} ratio of HF11T is 10.14× larger than the considered cells, as it has high on current (I_{on}) and low leakage current (I_{off}). RSNM of HF11T is 1.07× more than other compared cell due to the use of read decoupled structure. The results of various SRAM cells in 45 nm CMOS technology are complied in Table 2 at supply voltage of 0.6 V. From Table 2, it is evident that the proposed HF11T SRAM cell has superior performance compared to other single BL SRAM cells. As a result, the HF11T is an appropriate alternative for submicrometer technologies that require low power, high speed, and better stability.

References

[1] N. Maroof and B. S. Kong, "10T SRAM Using Half-VDD Precharge and Row-Wise Dynamically Powered Read Port for Low Switching Power and Ultralow RBL Leakage," IEEE Trans. Very Large Scale Integr. Syst., vol. 25, no. 4, pp. 1193–1203, 2017, doi: 10.1109/TVLSI.2016.2637918.

[2] S. Pal and A. Islam, "9-T SRAM Cell for Reliable Ultralow-Power Applications and Solving Multibit Soft-Error Issue," IEEE Trans. Device Mater. Reliab., vol. 16, no. 2, pp. 172–182, 2016, doi: 10.1109/TDMR.2016.2544780.

[3] B. Wang, T. Q. Nguyen, A. T. Do, J. Zhou, M. Je, and T. T. H. Kim, "Design of an ultra-low voltage 9T SRAM with equalized bitline leakage and CAM-assisted energy efficiency improvement," IEEE Trans. Circuits Syst. I Regul. Pap., vol. 62, no. 2, pp. 441–448, 2015, doi: 10.1109/TCSI.2014.2360760.

[4] S. Pal and A. Islam, "9-T SRAM Cell for Reliable Ultralow-Power Applications and Solving Multibit Soft-Error Issue," IEEE Trans. Device Mater. Reliab., vol. 16, no. 2, pp. 172–182, 2016, doi: 10.1109/TDMR.2016.2544780.

[5] Y. W. Chiu et al., "40 Nm Bit-Interleaving 12T Subthreshold Sram With Data-Aware Write-Assist," file///C/PhD DTU/SRAM Mem. Pap. cell Issues/SRC ref Pap. 0.2 V 32-Kb 10T SRAM With 41 nW Standby.pdf, vol. 61, no. 9, pp. 2578–2585, 2014, doi: 10.1109/TCSI.2014.2332267.

Just Accepted

[6] I. J. Chang, J. J. Kim, S. P. Park, and K. Roy, "A 32 kb 10T sub-threshold sram array with bit-interleaving and differential read scheme in 90 nm CMOS," IEEE J. Solid-State Circuits, vol. 44, no. 2, pp. 650–658, 2009, doi: 10.1109/JSSC.2008.2011972.
[7] M. H. Tu, J. Y. Lin, M. C. Tsai, S. J. Jou, and C. Te Chuang, "Single-ended subthreshold SRAM with asymmetrical write/read-assist," IEEE Trans. Circuits Syst. I Regul. Pap., vol. 57, no. 12, pp. 3039–3047, 2010, doi: 10.1109/TCSI.2010.2071690.
[8] M. H. Chang, Y. Te Chiu, S. L. Lai, and W. Hwang, "A 1kb 9T subthreshold SRAM with bit-interleaving scheme in 65nm CMOS," Proc. Int. Symp. Low Power Electron. Des., pp. 291–296, 2011, doi: 10.1109/ISLPED.2011.5993652.
[9] R. Saeidi, S. Member, M. Sharifkhani, and K. Hajsadeghi, "A Subthreshold Symmetric SRAM Cell," vol. 61, no. 1, pp. 26–30, 2014.
[10] A. D. Write-assist et al., "A 40 nm 512 kb Cross-Point 8 T Pipeline SRAM With Binary Word-Line Boosting Control , Ripple," vol. 61, no. 12, pp. 3416–3425, 2014.
[11] H. Fujiwara et al., "Mixed-Vth Read-port and Boosted WL Scheme for IoT Applications," in IEEE Asian Solid-State Circuits Conference, 2016, pp. 185–188. doi: 10.1109/ASSCC.2016.7844166.
[12] Y. H. Chen et al., "A 16nm 128Mb SRAM in high-? metal-gate FinFET technology with write-assist circuitry for low-VMIN applications," Dig. Tech. Pap. - IEEE Int. Solid-State Circuits Conf., vol. 57, pp. 238–239, 2014, doi: 10.1109/ISSCC.2014.6757416.
[13] I. Chatterjee et al., "Impact of technology scaling on SRAM soft error rates," IEEE Trans. Nucl. Sci., vol. 61, no. 6, pp. 3512–3518, 2014, doi: 10.1109/TNS.2014.2365546.
[14] V. Sharma, S. Vishvakarma, S. S. Chouhan, and K. Halonen, "A write-improved low-power 12T SRAM cell for wearable wireless sensor nodes," Int. J. Circuit Theory Appl., vol. 46, no. 12, pp. 2314–2333, 2018, doi: 10.1002/cta.2555.
[15] S. Pal, S. Bose, W. H. Ki, and A. Islam, "Reliable write assist low power SRAM cell for wireless sensor network applications," IET Circuits, Devices Syst., vol. 14, no. 2, pp. 137–147, 2020, doi: 10.1049/iet-cds.2019.0050.
[16] Y. W. Chiu et al., "40 Nm Bit-Interleaving 12T Subthreshold Sram With Data-Aware Write-Assist," file///C/PhD DTU/SRAM Mem. Pap. cell Issues/SRC ref Pap. 0.2 V 32-Kb 10T SRAM With 41 nW Standby.pdf, vol. 61, no. 9, pp. 2578–2585, 2014, doi: 10.1109/TCSI.2014.2332267.
[17] S. Pal, S. Bose, W. H. Ki, and A. Islam, "Half-Select-Free Low-Power Dynamic Loop-Cutting Write Assist SRAM Cell for Space Applications," IEEE Trans. Electron Devices, vol. 67, no. 1, pp. 80–89, 2020, doi: 10.1109/TED.2019.2952397.
[18] E. Abbasian and M. Gholipour, "Single-ended half-select disturb-free 11T static random access memory cell for reliable and low power applications," Int. J. Circuit Theory Appl., vol. 49, no. 4, pp. 970–989, 2021, doi: 10.1002/cta.2954.
[19] A. Sinha and A. Islam, "Low-power half-select free single-ended 10 transistor SRAM cell," Microsyst. Technol., vol. 23, no. 9, pp. 4133–4144, 2017, doi: 10.1007/s00542-016-3032-y.
[20] D. Shi, J. Yuan, J. Yin, Y. Wang, and S. Qiao, "A bit-interleaving 12T bitcell with built-in write-assist for sub-threshold SRAM," IEICE Electron. Express, vol. 19, no. 10, pp. 1–6, 2022, doi: 10.1587/elex.19.20220089.
[21] Y. He, J. Zhang, X. Wu, X. Si, S. Zhen, and B. Zhang, "A half-select disturb-free 11T SRAM cell with built-in write/read-assist scheme for ultralow-voltage operations," IEEE Trans. Very Large Scale Integr. Syst., vol. 27, no. 10, pp. 2344–2353, 2019, doi: 10.1109/TVLSI.2019.2919104.
[22] S. Ahmad, B. Iqbal, N. Alam, and M. Hasan, "Low Leakage Fully Half-Select-Free Robust SRAM Cells With BTI Reliability Analysis," IEEE Trans. Device Mater. Reliab., vol. 18, no. 3, pp. 337–349, 2018, doi: 10.1109/TDMR.2018.2839612.
[23] H. Kumar and V. K. Tomar, "Design of Low Power with Expanded Noise Margin Subthreshold 12T SRAM Cell for Ultra-Low Power Devices," J. Circuits, Syst. Comput., vol. 30, no. 6, 2021, doi: 10.1142/S0218126621501061.
[24] E. Shakouri, B. Ebrahimi, N. Eslami, and M. Chahardori, "Single-Ended 10T SRAM Cell with High Yield and Low Standby Power," Circuits, Syst. Signal Process., vol. 40, no. 7, pp. 3479–3499, 2021, doi: 10.1007/s00034-020-01636-y.
[25] L. Soni and N. Pandey, "A novel CNTFET based Schmitt-Trigger read decoupled 12T SRAM cell with high speed, low power, and high Ion/Ioff ratio," AEU - Int. J. Electron. Commun., vol. 167, no. April, 2023, doi: 10.1016/j.aeue.2023.154669.
[26] J. Singh, S. P. Mohanty, and D. K. Pradhan, Robust SRAM designs and analysis, 1st ed. New York: Springer, 2013. doi: 10.1007/9781461408185.
[27] S. Pal, S. Bose, W. H. Ki, and A. Islam, "A highly stable reliable SRAM cell design for low power applications," Microelectron. Reliab., vol. 105, no. August 2019, p. 113503, 2020, doi: 10.1016/j.microrel.2019.113503.
[28] M. Gupta, K. Gupta, and N. Pandey, "A novel PVT-variation-tolerant Schmitt-trigger-based 12T SRAM cell with improved write ability and high ION/IOFF ratio in sub-threshold region," Int. J. Circuit Theory Appl., vol. 49, no. 11, pp. 3789–3810, 2021, doi: 10.1002/cta.3134.
[29] B. Rawat and P. Mittal, "A 32 nm single-ended single-port 7T static random access memory for low power utilization," Semicond. Sci. Technol., vol. 36, no. 9, 2021, doi: 10.1088/1361-6641/ac07c8.
[30] E. Abbasian and M. Gholipour, "Improved read/write assist mechanism for 10-transistor static random access memory cell," Int. J. Circuit Theory Appl., vol. 50, no. 10, pp. 3642–3660, 2022, doi: 10.1002/cta.3361.
[31] N. Bai, X. Xiao, Y. Xu, Y. Wang, L. Wang, and X. Zhou, "Soft-Error-Aware SRAM With Multinode Upset Tolerance for Aerospace Applications," IEEE Trans. Very Large Scale Integr. Syst., vol. PP, pp. 1–9, 2023, doi: 10.1109/tvlsi.2023.3328717.
[32] P. Praveen and R. K. Singh, "Design of Enhanced Reversible 9T SRAM Design for the Reduction in Sub-threshold Leakage Current with14nm FinFET Technology," ACM Trans. Des. Autom. Electron. Syst., vol. 28, no. 6, pp. 1–29, 2023, doi: 10.1145/3616538.

Just Accepted

Received 10 September 2023; revised 11 February 2024; accepted 12 March 2024

Just Accepted

Electrical Engineering
https://doi.org/10.1007/s00202-024-02454-0

ORIGINAL PAPER

Adaptive control strategy for isolated renewable energy-based generation system with intermittency

Sombir Kundu[1] · Ashutosh K. Giri[2] · Madhusudan Singh[3] · Sunil Kadiyan[4]

Received: 27 November 2023 / Accepted: 29 April 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract
This paper implements a three-phase four-wire renewable generation system using wind and solar energy sources that feed standalone consumer loads. A solar photovoltaic (SPV) array, a bidirectional converter (BDC), and a battery storage system (BSS) are being integrated into a self-excited induction generator-based wind energy conversion system. A Modified Proportionate Affine Projection Algorithm (MPAPA) is designed to compute the amplitude of the fundamental load current component with reduced oscillations and keep the power equilibrium between the energy sources and load. An MPAPA-based control of voltage source inverter (VSI) provides numerous functions like harmonics abatement, compensation of neutral current, active and reactive power, load leveling, and voltage stabilization at the common point of intersection (CPI). A BDC controls the charging/draining of BSS by buck–boost modes operation and MPPT of the SPV during intermittent conditions. Moreover, the BSS is connected at the DC link of the VSI to absorb extra power and provide power support during intermittency. Moreover, a comparison is performed between the proposed MPAPA control and existing LMS and leaky momentum control approaches. The MPAPA control settles in less than one cycle (18 ms), becoming within the 2% specified limit. The dynamic simulation and test results of the isolated generation system are assessed under intermittency in non-conventional energy sources and unbalanced load. The total harmonic distortion of the CPI current is less than 5%, which meets the boundary of IEEE-519 standard.

Keywords Adaptive control approach · Battery storage system · Isolated power generation · Power quality

Abbreviations

3P4W, SEIG	Three-phase four-wire, self-excited induction generator
SPV, IC	Solar photovoltaic, incremental conductance,
BSS, VSC	Battery storage system, voltage source converter
WECS, CPI	Wind energy conversion system, common point of inter-section
MPAPA, THD	Modified Proportionate Affine Projection Algorithm, total harmonic distortion
BDC	Bidirectional converter
PLL, FLL	Phase-locked loop, frequency-locked loop
PQ, MLMS	Power quality, modified least mean square

✉ Sombir Kundu
sskundu46@gmail.com

Ashutosh K. Giri
ashu_ee2002@rediffmail.com

Madhusudan Singh
madhusudan@dce.ac.in

Sunil Kadiyan
sunkadiyan@gmail.com

[1] Department of Electrical Engineering, Ganga Institute of Technology and Management, Jhajjar, India

[2] Department of Electrical Engineering, Government Engineering College, Bharuch, Gujarat, India

[3] Department of Electrical Engineering, Delhi Technological University, Delhi, India

[4] Department of Mechanical Engineering, Ganga Institute of Technology and Management, Jhajjar, India

1 Introduction

Electricity usage has been rapidly growing owing to continuous increases in the global population. Global electricity demand is expanding twice as fast as primary energy demand, increasing electricity's proportion of total energy utilization from 19% in 2018 to 24% by 2040. Presently global

Published online: 13 June 2024

electricity demand is mainly met through fossil fuels-based generation. However, continuous depletion coupled with environmental problems due to the emission of greenhouse gases has attracted the interest of researchers toward the use of renewable energy sources. Because of this growing environmental concern, world is focusing on renewable energy-based electricity generation such as hydro, solar and wind energies among others. All these account for approximately 40% of the total installed energy capacity. With sustained cost reduction and advancement in technique, solar energy-based generation has increased significantly [1–4]. Hydropower plants even if they contribute less to the global energy supply have quite a large potential worldwide, and the development of hydropower for sustainable energy supply continued in the recent part but deployment rates accelerated for wind and SPV systems [5, 6]. However, out of all renewable resources wind energy stands out because of its high proficiency and low contamination. In recent years, continued technological advances in wind turbine technology have enhanced the strength of the WECS [7, 8]. Small-scale wind energy systems have been gaining popularity for off-grid or remote applications. Different electrical machines are taken into consideration for electromechanical conversion in small wind turbines, depending on the particular requirements and conditions of the application [9, 10].

The induction machine, and more specifically the three-phase squirrel induction machine, is still commonly employed for this peculiar operation. Indeed, the construction is sturdy, requires minimal servicing and is reasonably priced. As a result, substantial research has been conducted on SEIGs in remote and isolated places [11, 12]. Kundu et al. [13] and Sharma et al. [14] have identified significant problems with isolated renewable-based resources, such as variability and intermittency. As a result, although discussing the power quality (PQ) standard, it causes technical issues such as active power-sharing, unbalanced load adjustment, and reactive power compensation. Maintaining microgrid stability in islanded mode is a huge task. Voltage/frequency oscillations in the microgrid are minimized by using a BSS at the DC link to maintain the necessary power equilibrium. To adjust the characteristics of an isolated system under a variety of operating circumstances, an adaptive control strategy is utilized [15, 16]. Adaptive and a novel approach [17]-based control for WECS is described to work under intermittent conditions [18, 19]. A control approach for SPV panels to draw maximum power is discussed [20, 21]. To harness maximal power using INC and P&O MPPT techniques are described [22, 23] and an MLMS control is discussed to generate gating pulses for VSC [24, 25]. A hybrid microgrid system is developed and discusses the various challenges in microgrid systems [26]. Various PLL-based control schemes are available for PV-based generation. The development of generators in recent years has made wind energy an even better choice for a grid-connected or islanded DG system [27]. Controllers are required in wind turbines and solar photovoltaic systems to ensure low expenses of maintenance and effective operation [28]. The BSS is interfaced at the DC link to absorb the excess electricity [29], and the stored power in BSS is utilized during high load demand [30]. Saeed Golestan has given a review of different types of PLL (phase-locked loop) controllers with their pros and cons [31]. The control strategy for wind turbines and solar PV systems needs a PLL controller based on an all-pass filter (APF) to mitigate PQ issues in a standalone generation. PLL cannot produce unit vectors by itself, and the inclusion of a moving average filter is utilized in the loop which leads to a sluggish response. As a result of the improved PQ in the DG system, PLL based on APF is recommended [32]. The APF allows all types of frequencies to pass through without making any changes in their magnitude. It is only responsible for changing its phase change. In single-phase systems, the phase shift is required to generate a fictitious orthogonal signal, while in three-phase systems the phase shift is required to cancel the unbalances of grid voltage [33]. The controls are employed to limit the PQ issues of a wind–SPV–battery-based microgrid under transient load conditions [34]. Further, the control algorithms are employed to provide synchronization and smooth control of the standalone system for remote applications [35]. In the present work, an adaptive MPAPA control approach [36, 37] is designed to manage the voltage profile and frequency in a wind–solar PV–BSS-based hybrid system for all operating conditions. This controller yields rapid response and is acceptable for applications of rapid dynamics like wind and solar energy-based power generation.

The authors' originality and substantial contribution are listed as:

(1) An improved filtering strategy based on adaptive MPAPA is performed to estimate the amplitude of fundamental component to preserve sinusoidal and stabilized source currents even under the irregular conditions of solar radiations, wind speed, and polluted load.
(2) The proposed MPAPA control with VSI yields a fast response in the regulation of frequency and voltage, compensation of neutral load current and reactive power, harmonics eradication, balancing of load and reduces the PQ issues of the standalone system.
(3) Under different dynamic circumstances, the BSS response is examined, and the total harmonic distortion (THD) of CPI current is examined based on simulation and hardware analysis which are well below the IEEE-519 permissible level.
4) Moreover, a comparison is made between the proposed adaptive MPAPA control and the existing control techniques. The settling time of the proposed adaptive

MPAPA control is found to be shorter than one cycle (18 ms) within the 2% specified limit.

2 System description

The major components of the 3P4W standalone DG system are a 3-phase SEIG, wind turbine, solar PV array, IGBT-based VSI, and BSS. The capacitor bank is being utilized to provide the desired reactive power at no load for the SEIG. The capacitor bank serves dual functions: First, it reduces the high noise generated by the controllers, and second, it provides reactive power support to the proposed generator. All other device ratings are selected with proper design considerations. The wind turbine acts as a prime mover for the proposed SEIG, which in turn helps the generator's rotor to turn faster than synchronous speed, thus acting as an induction generator [13]. The generated energy is supplied to the three-phase nonlinear load. Here, we have also used an IC-based MPPT technique to extract maximal power from an SPV. Further, a BSS is interfaced at a DC link with a BDC controller to consume the excess electricity and supply power during intermittency and increase in consumer [14].

3 Control strategy

Figure 1 represents the control mechanism of the 3P4W standalone renewable-based isolated system (Fig. 2). An adaptive MPAPA control is designed to estimate the fundamental component of the polluted and unbalanced load current under the uncertain condition of NCES's. Further, the proposed control produces the triggering pulses for the VSI as depicted in Fig. 3.

3.1 Maximum power extraction and control scheme of bidirectional controller

The DC-link voltage oscillations are caused due to real power imbalances. Active power imbalances/disruptions can occur for a variety of reasons, including (1) deviation in solar power generation, (2) changes in wind speed power generation, and (3) unbalanced and increased in consumer loads. The BDC controller adjusts the DC-link voltage and controls the BSS charging and draining modes in intermittent wind speed and solar radiations. The proposed BDC controller is constructed using two PI controllers (inner and outer) for BSS management. The outer controller is in charge of managing the DC-link voltage and tracking the PV array's maximum power. The inner side controller loop is employed to trace the reference current of the BSS. The outer PI controller employed to control the DC-link voltage is denoted as

$$i_{\mathrm{b}}(q) = i_{\mathrm{b}}(q-1) + k_{\mathrm{bp}}\{v_{\mathrm{dce}}(q) - v_{\mathrm{dce}}(q-1)\} + k_{\mathrm{bi}}(q) \tag{1}$$

$$\text{where,}\quad v_{\mathrm{dce}}(q) = v_{\mathrm{mpp}}(q-1) - v_{\mathrm{dc}}(q) \tag{2}$$

where v_{mpp} is the evaluated reference voltage at the DC link from an IC-MPPT algorithm. The actual DC-link voltage is denoted by V_{dc}. The duty cycle (α) of the BDC is described as the outcome of the inner PI controller.

$$\alpha(q) = \alpha(q-1) + k_{\mathrm{sp}}\{i_{\mathrm{be}}(q) - i_{\mathrm{be}}\,(q-1)\} + k_{\mathrm{si}} i_{\mathrm{be}}(q) \tag{3}$$

where (i_{be}) is the difference between the reference battery current (i_{b}^{*}) and the observed battery current (i_{b}) [16].

$$i_{\mathrm{be}}(q) = i_{\mathrm{b}}^{*}(q) - i_{\mathrm{b}}(q) \tag{4}$$

3.2 Control strategy for VSI

Primarily, the structure of the MPAPA is established, which serves as the foundation for our subsequent derivations. Let (μ_{opt}) denote the plant's unknown impulse response vector and ($\xi^{(i)}$) the adaptive filter's input signal vector as displayed in Fig. 2 [37].

$$\mu_{\mathrm{opt}}^{(i)} = \left[\mu_0^{(i)}, \mu_1^{(i)}, \ldots\ldots \mu_{M-1}^{(i)}\right]^T \tag{5}$$

$$\xi^{(i)} = [\xi^{(i)},\ \xi^{(i-1)}, \ldots \xi^{(i-k+1)}] \tag{6}$$

where (M) signifies the adaptive filter length and (i) indicates the current sample. The recorded signal ($X^{(i)}$) with added impulsive noise ($N^{(i)}$) can be computed from Fig. 2 as (7)

$$X^{(i)} = \mu_{\mathrm{opt}}(\xi^{(i-1)})^T + N^{(i)} \tag{7}$$

The output of the undetermined system is subjected to the calculated posterior, and the adaptive filter is

$$\mu_{\mathrm{opt}} = X^{(i)} - (\xi^{(i)})^T \xi^{(i)} \tag{8}$$

After each iteration, $\xi^{(i)}$ is computed using $\mu_{\mathrm{opt}}^{(i)}$. It is possible to calculate the novel MPAPA update weight coefficient as

$$\mu^{(i)} = \mu^{(i-1)} + \rho\xi^{(i)} X^{(i)}\left[(\xi^{(i)})^T \xi^{(i)} X^{(i)} + \delta I\right]^{-1} \eta^{(i)} \tag{9}$$

The gain-controlling matrix is denoted by $X^{(i)}$=diag[$x_0^{(i)}$, $x_1^{(i)}$, $\ldots..x_{M-1}^{(i)}$], where I is an (M × M) identity matrix

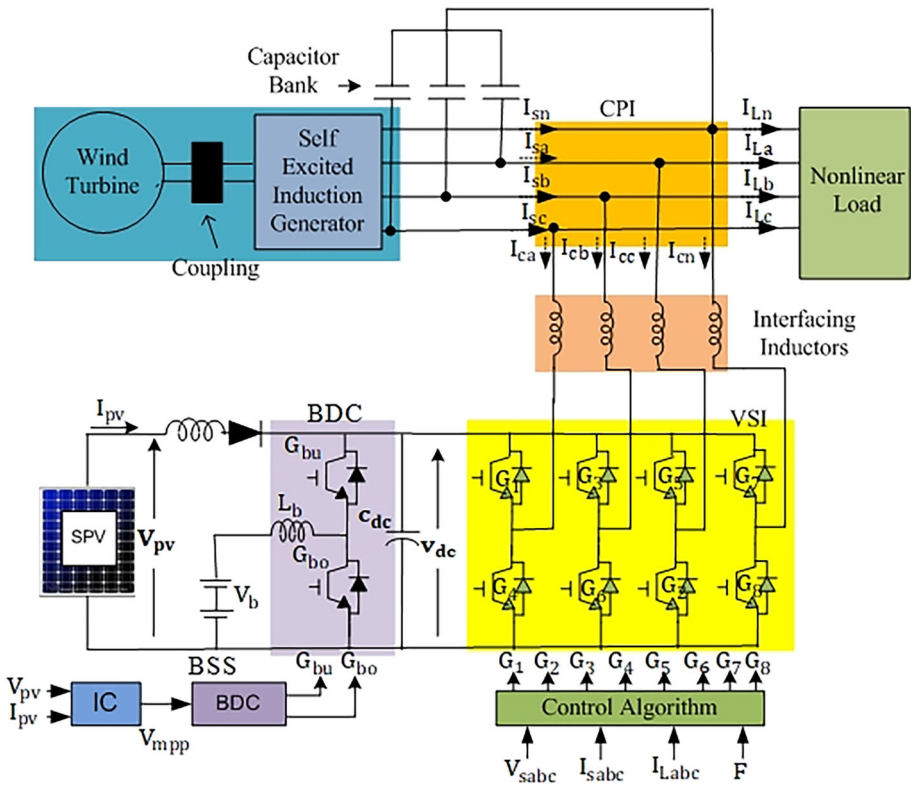


Fig. 1 Control topology of 3P4W standalone system

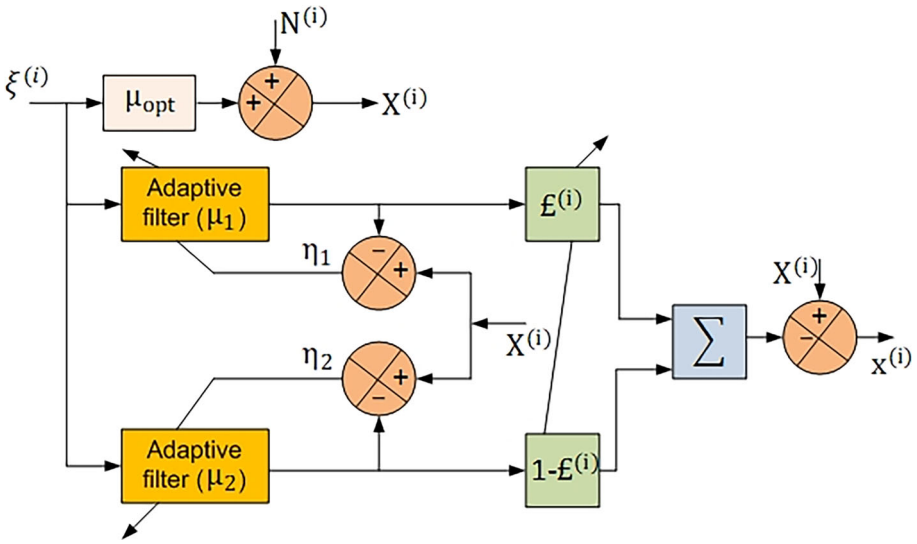


Fig. 2 Schematic diagram of MPAPA control

and (ρ) is the iteration's step size. (10), which represents the individual gain unit ($x_k^{(i-1)}$, $k \in [0, \ldots, M-1)$.

$$x_k^{(i-1)} = \frac{M d_k^{(i-1)}}{\sum_{k=0}^{M-1} d_k^{(i-1)}} \tag{10}$$

In this case, $d_k^{(i-1)} = \max\{\mathrm{p} \max\{\mathrm{q} \left|\mu_{\mathrm{opt}}^{(i)}\right|\}, \left|\mu_k^{(i-1)}\right|\}$. The initialization technique has to employ the parameter p when the procedure begins with all weight coefficients set to zero. This concludes the weight coefficients from stopping during updates. Similarly, although having a considerably lower value than the maximum coefficient, the parameter (q) is a relatively positive quantity that eliminates ($\mu_{M-1}^{(i)}$) from stalling.

Throughout the iteration process, the largest step size (ρ_1) and the smallest step size (ρ_2) get recorded and meet the requirement $0 \leq \rho_2 \leq \rho_1$.

Taking into consideration these two distinct step sizes, given in (11) & (12), the value of the weight coefficient $\mu^{(i-1)}$

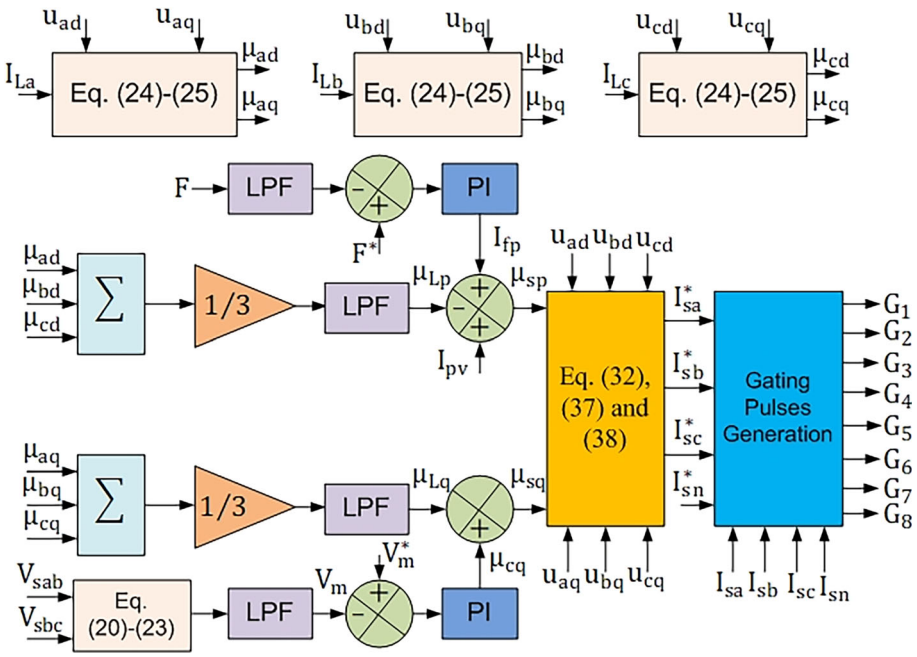


Fig. 3 Control strategy of the 3P4W standalone system

is adjusted as $\mu_1^{(i)}$ and $\mu_2^{(i)}$.

$$\mu^{(i)} = \mu^{(i-1)} + \rho_1 \xi^{(i)} X^{(i)} [(\xi^{(i)})^T \xi^{(i)} X^{(i)} + \delta I]^{-1} \eta^{(i)} \quad (11)$$

$$\mu^{(i)} = \mu^{(i-1)} + \rho_2 \xi^{(i)} X^{(i)} [(\xi^{(i)})^T \xi^{(i)} X^{(i)} + \delta I]^{-1} \eta^{(i)} \quad (12)$$

With the restriction $0 \leq £^{(i)} \leq 1$, the two new weight vectors, $\mu_1^{(i)}$ and $\mu_2^{(i)}$, are arranged by a variable mixing factor $£^{(i)}$ to update the weight vector as (13).

$$\mu^{(i)} = £\mu_1^{(i)} + (1 - £^{(i)}) £ \mu_2^{(i)} \quad (13)$$

Substituting (11) & (12) into (13), the weight vector coefficient's newly determined equation is (14).

$$\lambda^{(i)} = £^{(i)} \lambda_1 + (1 - £^{(i)}) \lambda_2 \quad (14)$$

The total number of steps size, denoted as $\lambda^{(i)}$, ranges from λ_1 to λ_2. A sigmoid activation function's output concerning a variable $\beta^{(i)}$ is the definition of the variable mixing factor ρ, which may be obtained as (15).

$$\rho^{(i)} = \text{sgn}\left[\beta^{(i)}\right] = \frac{1}{1 + e^{-\beta^{(i)}}} \quad (15)$$

In this case, ρ is determined at each iteration by minimizing the system output error L_1-norm.

$$Y^{(i)} = X^{(i)} - (\xi^{(i)})^T \mu^{(i-1)} \quad (16)$$

$\beta^{(i)}$ has been updated using the gradient descent method to optimize the L_1-norm of the system error $Y^{(i)} = X^{(i)} - (\xi^{(i)})^T \mu^{(i-1)}$, ensuring the algorithm's robustness in the existence of non-Gaussian impulsive interference. Equation (17) can be obtained by integrating Eqs. (14) and (15) to obtain the updated $\mu^{(i)}$.

$$\mu^{(i)} = \mu^{(i-1)} + \frac{\lambda \xi^{(i)} X^{(i)} \eta^{(i)}}{\left[(\xi^{(i)})^T \xi^{(i)} X^i + \delta I\right]} \quad (17)$$

The step size of the adaptation $(\beta^{(i)})$ is represented by (ρ_β) in this case, and D_1 is a positive value that keeps the simulation from halting when $£^{(i)}$ becomes 0 or 1. The objective of the MPAPA approach is to determine, from the input signal $\xi^{(i)}$ and the preferred output signal $x(i)$, the genuine system impulse response $\mu^{(i)}$. Since the sigmoid activation function's output (18) defines the variable mixing factor, it is limited to falling inside the open interval (0, 1).

$$\beta^{(i)} = \beta^{(i-1)} - \rho_\beta \frac{\partial |\eta^{(i)}|}{\partial |\beta^{(i-1)}|} \quad (18)$$

The revised $\rho^{(i)}$ can now be expressed as

$$\rho^{(i)} = \begin{Bmatrix} 1, & \text{if} > -\ln \frac{D+1}{D-1} \\ 0, & \text{if} < -\ln \frac{D+1}{D-1} \\ \frac{D}{1+e^{\rho(i)}} - \frac{D}{2} + 0.5, & \text{otherwise} \end{Bmatrix} \quad (19)$$

where $\rho^{(i)} = \text{sgn}[\beta^{(i)}]$. It is possible to modify (19) by adding a modest positive integer to keep the simulation from halting when $\rho^{(i)}$ is equal to 1 or 0.

3.3 Estimating amplitude of supply voltage and unit template generation for reference current

V_m Is the voltage amplitude of the terminal voltage, which is expressed as

$$V_m = \left(\frac{2}{3}\right)^{1/2}\left(v_{sa}^2 + v_{sb}^2 + v_{sc}^2\right)^{1/2} \tag{20}$$

Here, phase voltages are estimated as

$$\begin{aligned} v_{sa} &= \frac{1}{3}\left(2v_{sab} + v_{sbc}\right), \quad v_{sb} \\ &= \frac{1}{3}\left(-v_{sab} + v_{sbc}\right), \quad v_{sc} = \frac{1}{3}\left(-v_{sab} + 2v_{sbc}\right) \end{aligned} \tag{21}$$

Thus, by using Eqs. (20, 21), the unit templates (in-phase and quadrature) are evaluated as [9]

$$u_{ad} = v_{sa}/V_m, \quad u_{bd} = v_{sb}/V_m, \quad u_{cd} = v_{sc}/V_m \tag{22}$$

$$\begin{aligned} u_{aq} &= -\frac{u_{ad}}{\sqrt{3}} + \frac{u_{cd}}{\sqrt{3}}, \quad u_{bq} = -\frac{3u_{cd}}{2\sqrt{3}} - \frac{u_{cd}}{2\sqrt{3}} + \frac{u_{bd}}{2\sqrt{3}}, \\ u_{cq} &= -\frac{3u_{ad}}{2\sqrt{3}} + \frac{u_{bd}}{2\sqrt{3}} - \frac{u_{cd}}{2\sqrt{3}} \end{aligned} \tag{23}$$

The new weight in (17) is converted to weighted values corresponding to the active power components (24), and the reactive components are formed as (25).

$$\mu_{px}^{(i)} = \mu_{px}^{(i-1)} + \frac{\left[£^{(i)}\rho_1 + (1 - £^{(i)})\rho_2\right]}{\left[\left(\xi_{px}^{(i)}\right)X^{(i)}{\xi_{px}}^{(i)} + \delta I\right]^{-1}}\eta_x^{(i)} \tag{24}$$

$$\mu_{qx}^{(i)} = \mu_{qx}^{(i-1)} + \frac{\left[£^{(i)}\rho_1 + (1 - £^{(i)})\rho_2\right]}{\left[\left(\xi_{qx}^{(i)}\right)X^{(i)}{\xi_{qx}}^{(i)} + \delta I\right]^{-1}}\eta_x^{(i)} \tag{25}$$

From Eq. (26) comes the average per phase fundamental active weight component (μ_{lp}). Equation (27) can also be used to compute the reactive weight component (μ_{lq}).

$$\mu_{lp} = \frac{1}{3}\sum_{x\epsilon\{a,b,c\}} \mu_{px} \tag{26}$$

$$\mu_{lq} = \frac{1}{3}\sum_{x\epsilon\{a,b,c\}} \mu_{pq} \tag{27}$$

Further, the frequency error is computed by subtracting the reference frequency (50 Hz) from the system frequency as below

$$F_e(q) = F^*(q) - F(q) \tag{28}$$

The system frequency is computed using a phase-locked loop (PLL). The frequency-side PI controller output (i_{fp}) is computed as

$$i_{fp}(q) = i_{fp}(q-1) + k_{pf}\{F_e(q) - F_e(q-1)\} + k_{if}F_e(q) \tag{29}$$

Therefore, to evaluate the total fundamental active current (μ_{sp}), the fundamental average load current output (μ_{Lp}) is subtracted from the addition of the output of the frequency PI controller (i_{fp}) and solar PV current (I_{pv}), and the equation is written as

$$\mu_{sd} = i_{fp} + I_{\text{pv}} - \mu_{Lp} \tag{30}$$

$$\text{where,} \quad \mu_{Lp} = \frac{\mu_{ad} + \mu_{bd} + \mu_{cd}}{3} \tag{31}$$

The active component of the reference current is calculated as (32).

$$i_{xd}^* = \mu_{sd}u_{ad}, \quad x \in \{a, b, c\} \tag{32}$$

The load current component, i.e., fundamental average active component per phase, is estimated using all three-phase active components in Eq. (31). After computing the extracted current (I_{Labc}) and active components (i_{xd}^*), the sample and hold logic is employed. As a result, one can find the active load current component (μ_{ad}) for phase 'a', and the (μ_{bd}) and (μ_{cd}) are the active load current components for phases '*b*' and '*c*'. The reactive component (μ_{cq}) is formed by the voltage PI controller. This is expressed mathematically as

$$\mu_{cq}(q) = \mu_{cq}(q-1) + k_{pq}\{v_e(q) - v_e(q-1)\} + k_{iq}\{v_e(q)\} \tag{33}$$

$$\text{where,} \quad V_{\text{merr}}(q) = V_m^*(q) - V_m(q) \tag{34}$$

The amplitude of terminal voltage (V_m) and its reference value at CPI is taken as (V_m^*). So, the total fundamental reactive component is evaluated as

$$\mu_{sq} = \mu_{cq} - \mu_{Lq} \tag{35}$$

$$\mu_{Lq} = \frac{\mu_{aq} + \mu_{bq} + \mu_{cq}}{3} \tag{36}$$

The reactive component of the reference current is determined as (37).

$$i_{xq}^* = \mu_{sq}u_{aq}, \quad x \in \{a, b, c\} \tag{37}$$

Therefore, the source reference currents are calculated as [36, 37] (38).

$$i_{sx}^{*} = i_{xd}^{*} + i_{xq}^{*} \quad x \in \{a, b, c\} \tag{38}$$

The gating pulses of the VSI are formed by the hysteresis current controller (HCC) [25]. However, the HCC approach is faster despite various limitations such as greater switching frequencies and changeable switching frequencies. The hysteresis band ($\delta = 0.01$) is chosen for the proposed controller's proper operation. The error is calculated by subtracting the detected grid currents (I_{sa}, I_{sb}, and I_{sc}) from the grid reference currents (I_{sa}^{*}, I_{sb}^{*}, and I_{sc}^{*}).

4 Results and discussion

The result of the proposed 3P4W standalone system is analyzed under intermittency in solar radiations, wind speed, and consumer load. During variations in load demand and changes in solar PV insolation, the reliability of the proposed MPAPA approach is confirmed through simulation and test results. The simulation results of CPI voltage (V_{sabc}), CPI current (I_{sabc}), load current (I_{Labc}), compensator current (I_{cabc}), DC-link voltage (V_{dc}), frequency (F), neutral load current (I_{Ln}), PV current (I_{pv}), duty cycle (α), and PV power (P) are taken under various operating circumstances.

4.1 Dynamic response of wind and SPV system during unbalanced load

Figure 4 indicates the dynamic response of the suggested wind and SPV system during constant wind speed/solar radiations feeding an unbalanced load. The solar radiations and wind speed are fixed at 900 w/m^2 and 15 m/s, respectively. The load of phase 'b' is detached at t = 2.7 s to t = 2.9 s. During the removal of a load of phase 'b', the load current is decreased and the controller manages the CPI voltage, CPI current, frequency, and DC-link voltage of the proposed DG system. The total generation of the standalone system is increased from 5.2 to 5.8 kW during the elimination of the load of phase 'b'. Further, the compensator current and neutral load current are increased during load detached. The MPAPA control with VSI provides multi-functions like voltage/frequency control, neutral current compensation, harmonics eradication, power support and enhances the overall PQ of the isolated system.

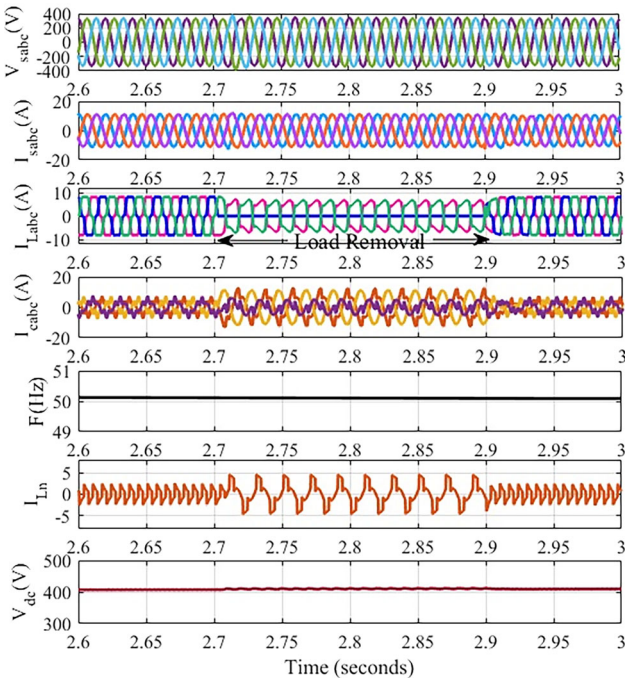


Fig. 4 Dynamic response of wind and SPV system during unbalanced load

4.2 Dynamic response under intermittency

The dynamic performance of the standalone wind and SPV system during wind variation feeding balanced nonlinear load is observed and displayed in Fig. 5a and b. Figure 5a represents the wind speed is altered from 15 to 12 m/s at t = 2.8 s and the solar radiation is shifted 500–900 w/m^2 at t = 2.7 s. Due to inflation in solar radiations, the PV current and the PV power of the proposed isolated system are increased from 5.9 to 11.77A and 1.8–2.8 kW, respectively, and the duty cycle is decreased. Further, the total power generation is reduced from 5.8 to 4.4 kW due to a decrease in wind speed from 15 to 12 m/s. Figure 5b depicts that during the variation in solar radiations at t = 2.7 s, the CPI voltage shows a slight deflection of 0.01 s, and then, the controller manages the CPI voltage. The (I_{sabc}) and (I_{Labc}) are increased during the intermittency of wind speed and solar radiation. Further, the VSI continuously provides power support during the change in wind/solar radiations with the help of a DC-link connected BSS. Moreover, the controller manages the (F) and (V_{dc}) of the renewable energy-based system.

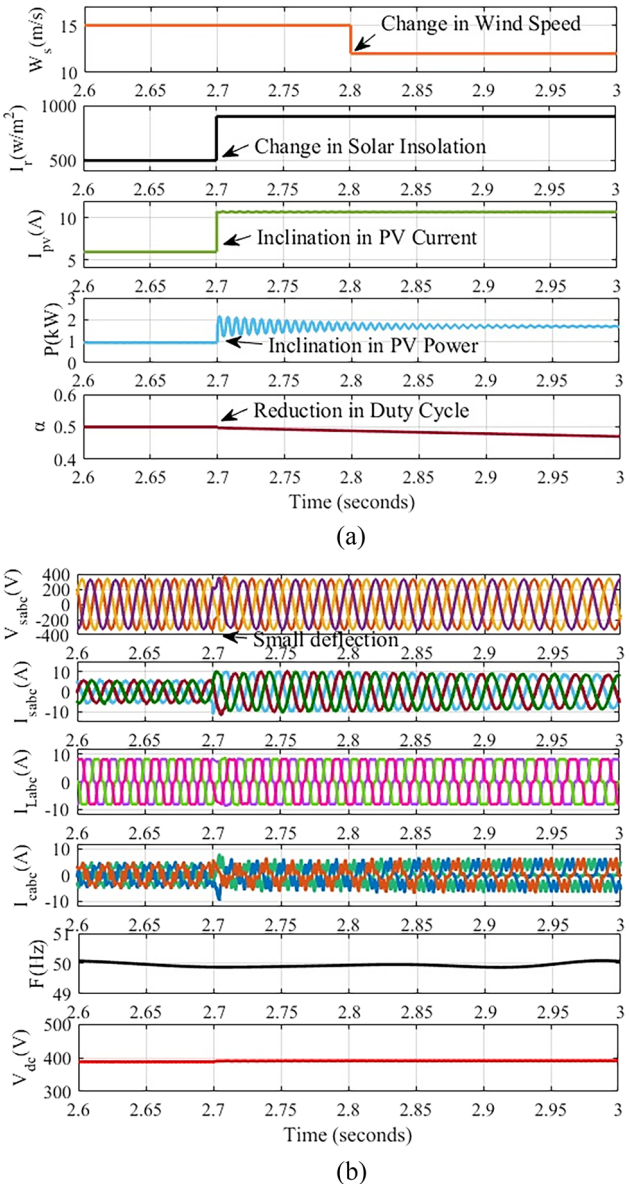


Fig. 5 **a** and **b** Dynamic response of renewable energy-based system under intermittency

4.3 Harmonic profiles of standalone system

Figure 6a, b illustrates the harmonic profiles of CPI current and load current of the proposed renewable-based isolated system. Figure 6a, b illustrates the THDs of (I_{sb}) and (I_{Lb}) of phase 'b' are 4.16% and 25.19%, respectively. The harmonic analysis of the 3P4W standalone system is analyzed under uncertain load conditions. The THDs of the standalone system are found in limit and follow the IEEE-514 standard.

5 Test result

The experimental configuration used to assess the performance of the MPAPA control approach under dynamic load is displayed in Fig. 7. An oscilloscope, dSPACE-1104 control board, voltage and current sensors, host computer, and DC power supply, power analyzer, nonlinear load make up the setup. Specifically, integration between the dSPACE-1104 control desk and the MATLAB/Simulink environment, where the proposed control is modeled, is made possible by the host computer. The MPAPA-based models are automatically generated code and then assembled and uploaded onto the dSPACE control board, allowing the dSPACE control desk to track the MPAPA estimations in real time. Lastly, the dSPACE digital-to-analog converter transmits the estimates to the oscilloscope.

5.1 Response Isolated system at unbalanced load

Figure 8a–f represents the steady-state and dynamic response of the proposed isolated system under various loading circumstances. The said figures show different signals of isolated system such as in-phase unit template (u_{ad}), CPI voltage (V_{sa}), CPI current (I_{sa}), load current (I_{La}), and DC-link voltage (V_{dc}). Figure 8a, b depicts the steady-state and dynamic response of u_{ad}, V_{sa}, I_{La}, and I_{sa} under nonlinear load conditions. Figure 8c shows the dynamic response of V_{sa}, I_{La}, $\mathrm{I_{sa}}$, and I_{sc} under load insertion of different phases at different times. During the insertion of load at different times as depicted in Fig. 8c, the I_{La}, I_{sa}, and I_{sc} are increased. Figure 8d, e shows the response under load insertion and load removal conditions. Figure 8d shows during the reduction in load, the I_{La} and I_{sa} are increased, and on another side in Fig. 8e shows the signals under insertion of load, the I_{La} and I_{sa} are reduced. Figure 8f reveals the performance of V_{dc}, V_{sa}, I_{sa}, and I_{La} during dynamics in load. The V_{dc}, I_{sa}, and I_{La} are decreased during the load reduction duration. The dynamic results illustrate the proposed MPAPA control approach is able to maintain the aforementioned parameters under unbalanced and highly nonlinear load conditions. Further, the CPI current of the proposed system remains sinusoidal during the exclusion or insertion of load.

5.2 BSS Response during intermittency

The transient response of BSS under various operating conditions is depicted in Fig. 9. The BSS response is analyzed: (i) uniform solar radiations/wind speed feeding balanced nonlinear load, (ii) solar radiations/constant wind speed under uncertain conditions of nonlinear load, (iii) change in wind speed and uniform solar radiations feeding balanced load, (iv) change in solar radiations and constant wind speed feeding

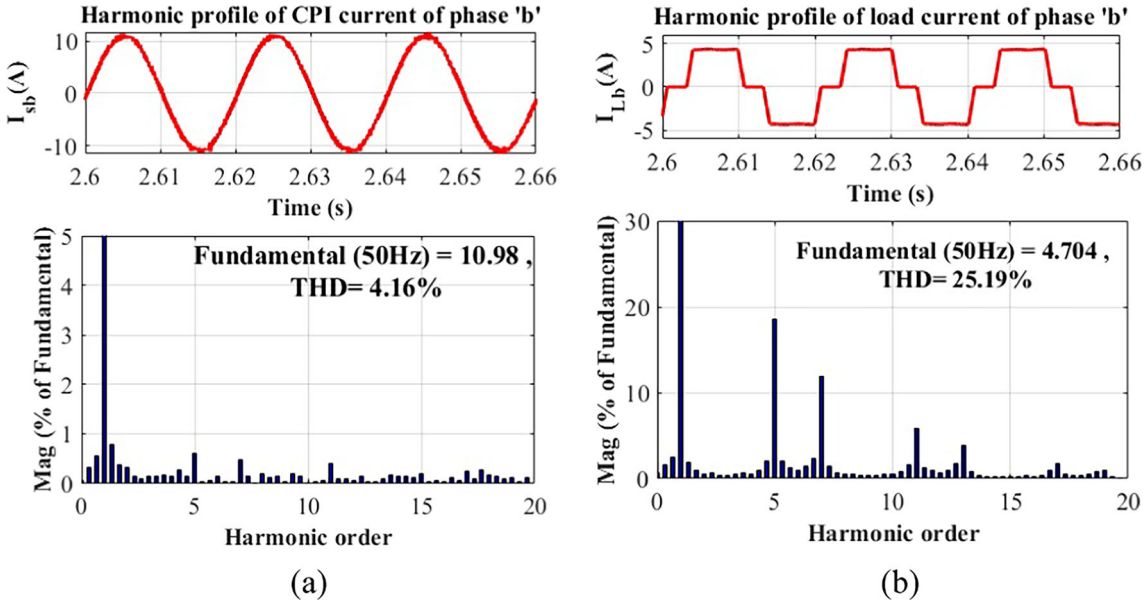


Fig. 6 **a** & **b** Harmonic profiles of the isolated system

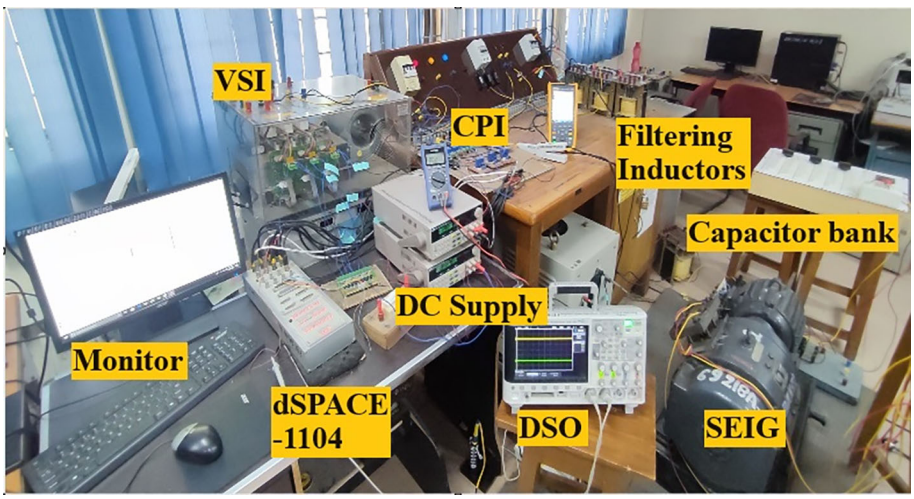


Fig. 7 Experimental prototype of isolated system

balanced load. During the elimination of a load of phase 'b', the excess power is transferred to the BSS. The BSS is shifted to the charging mode, and the BSS current is improved due to the removal of load. During the abrupt change in wind speed from 15 to 12 m/s, the BSS charging current is diminished. Also, during the decrement in solar radiations 900–500 w/m^2 at $t = 2.8$ s, the BSS is switched to draining mode and delivers power to CPI through VSI. The designed controller successfully balances all standalone system parameters under all operating situations.

Further, Table 1 depicts the comparison performance of proposed adaptive MPAPA, leaky momentum, and LMS-based existing control approaches. The parameters of the proposed standalone system are analyzed under different operating conditions and found that the proposed MPAPA control gives better performance, fast convergence, least oscillations with better stability, and minimum change in DC-link voltage.

5.3 Harmonic profiles of isolated system

Figure 10a, b represents the experimental results of harmonics under a polluted load. The THDs of CPI current and load current are 4.40% and 25.32%, respectively. The harmonic analysis of the 3P4W standalone system is analyzed under uncertain and highly nonlinear load conditions. On the observation of simulation and hardware results, the harmonic profile of the standalone system is found almost similar. The THDs of the standalone system are found in the limit and meet the IEEE-514 standard.

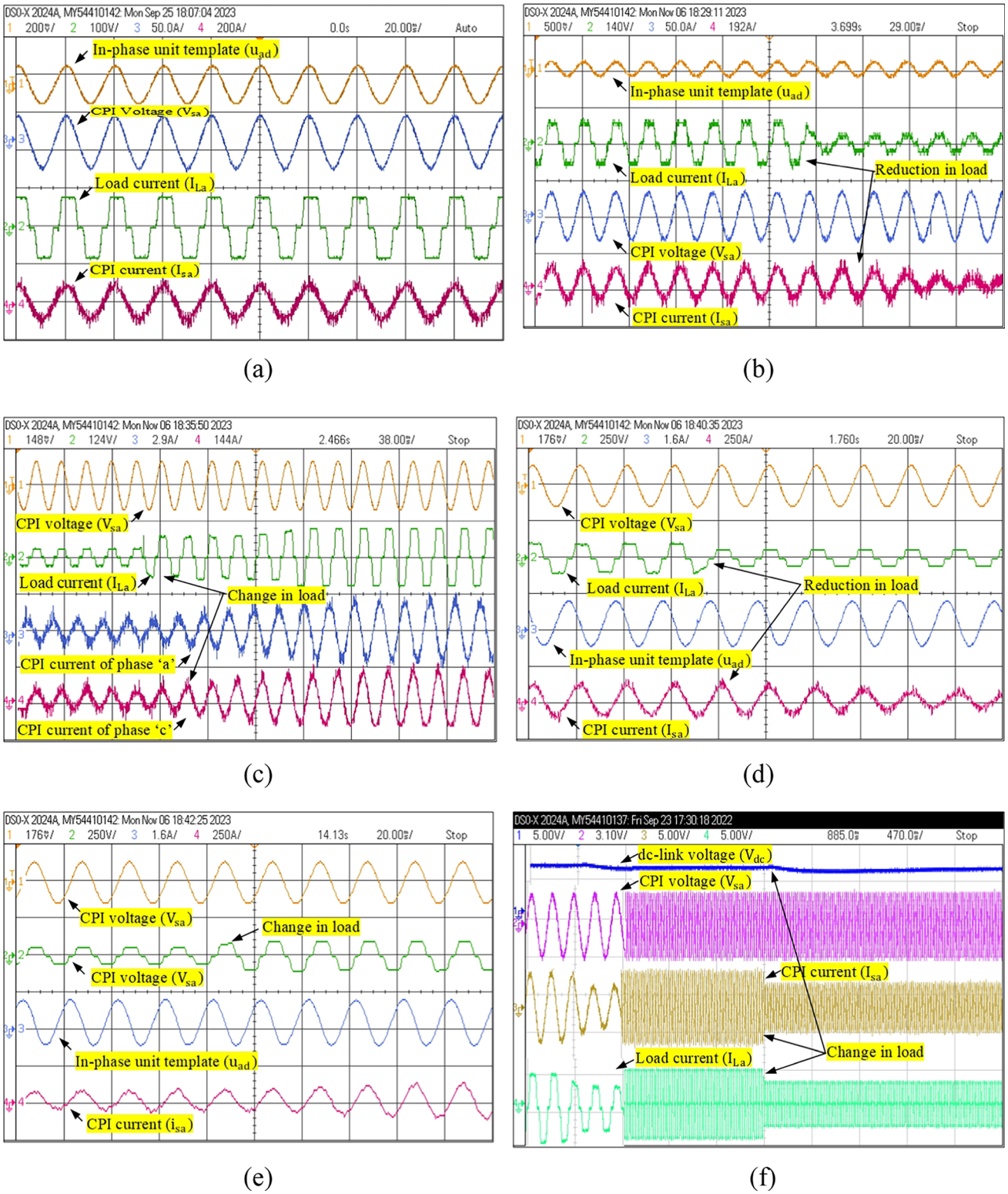


Fig. 8 Response of proposed isolated system at unbalanced load **a**, **b** u_{ad}, V_{sa}, I_{La} and I_{sa}, **c** V_{sa}, I_{La}, I_{sa}, and I_{sc} **d–e** V_{sa}, I_{La}, u_{ad}, and I_{sa} and **f** V_{dc}, V_{sa}, I_{sa}, and I_{La}

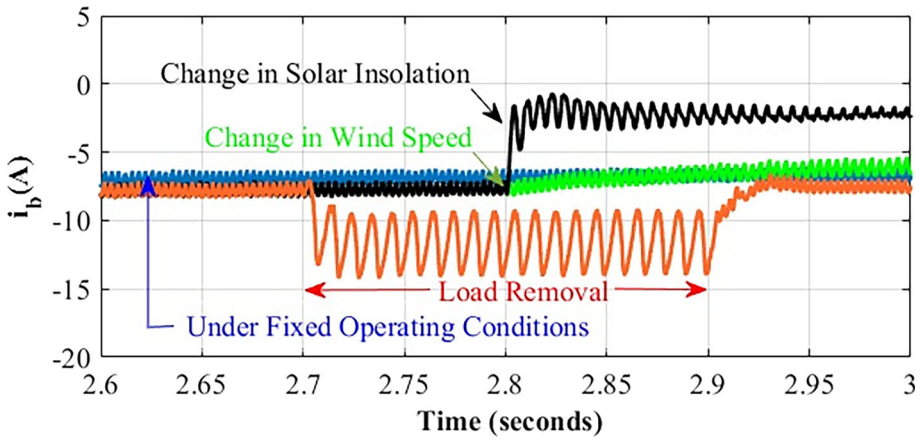
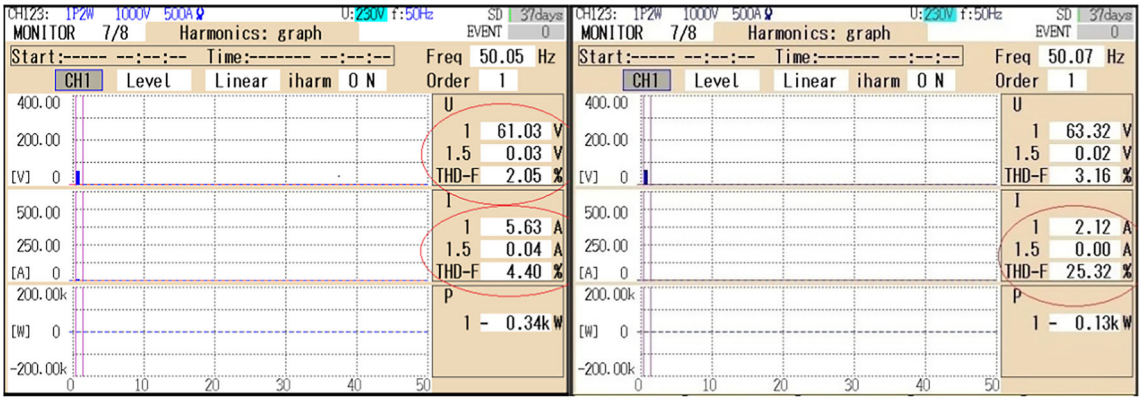


Fig. 9 BSS response during intermittency

Table 1 Comparison performance of the proposed MPAPA with exiting LMS and leaky momentum control approaches

S. No.	Parameters	Leaky momentum	LMS	Proposed MPAPA
1	Sampling time (T_s)	50 μs	50 μs	50 μs
2	THD% of (I_{sb})	4.8%	4.56%	4.40%
3	Change in DC-link voltage	4 V	4–5.5 V	3 V
4	Convergence	Medium	Medium	Least
5	Complexity	Medium	Medium	Medium



(a) (b)

Fig. 10 **a** & **b** Harmonic profiles of a standalone system

6 Conclusion

An in-loop filter-based adaptive MPAPA control algorithm is implemented to enhance resilience in wind and SPV-based standalone generation system with intermittency and testing consumer loads. A BSS of adequate capacity is being used to ensure power balance in this isolated system. The proposed MPAPA control successfully computed the reference source currents under intermittent conditions of wind speed, SPV radiations, and unbalanced loading. The simulation results indicated PQ enhancement with the stabilization of the frequency and voltage profile of the renewable-based standalone system. Moreover, the MPAPA control approach with VSI effectively provides power support, neutral current and reactive power compensation, and load leveling under all operating conditions. The simulation and hardware results confirmed that the CPI voltage and current THDs consistently lie within the IEEE-519-2014 acceptable limit. The elimination of the boost converter from the SPV stage did not affect MPP's performance. Furthermore, the MPP operation was accomplished using the BDC at all radiation levels. As a result, the proposed 3P4W standalone energy system with the MPAPA approach is a feasible alternative for rural electrification.

Author contributions 1) Author 1st has prepared the manuscript. 2) Authors 2nd and 3rd have reviewed the manuscript and help to design the hardware prototype. 3) Author 4th has help to taken the hardware result.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Shaikh MRS (2017) A review paper on electricity generation from solar energy. Int J Res Appl Sci Eng Technol 5(9):1884–1889. https://doi.org/10.22214/ijraset.2017.9272
2. Alwal LA, Kihato PK, Kamau SI (2018) A review of control strategies for microgrid with PV-wind hybrid generation systems. In: Proceedings of the sustainable research and innovation conference, pp. 243–250
3. Narayanan V, Singh B (2023) Seamless operation of an SRG based WEGS and a bidirectional converter interfaced BES microgrid under polluted grid conditions. IEEE Trans Ind Appl. https://doi.org/10.1109/TIA.2023.3306741
4. Kundu S, Singh M, Giri AK, Kadiyan S, Chittora P, Shailly D (2024) Robust and fast control approach for islanded microgrid system and EV charging station applications. Electr Eng. https://doi.org/10.1007/s00202-024-02291-1
5. Bagher AM, Vahid M, Mohsen M, Parvin D (2015) Hydroelectric energy advantages and disadvantages. Am J Energy Sci 2(2):17–20
6. Kewat S, Singh B (2020) Improved reweighted zero-attracting quaternion-valued LMS algorithm for islanded distributed generation system at im load. IEEE Trans Ind Electron 67(5):3705–3716. https://doi.org/10.1109/TIE.2019.2916301
7. Jain A, Saravanakumar R, Shankar S (2022) Moving average filter-PLL-based voltage and frequency control of standalone WECS. IETE J Res 68(5):3628–3637. https://doi.org/10.1080/03772063.2020.1772128
8. Chauhan PJ, Chatterjee JK (2019) A novel speed adaptive stator current compensator for voltage and frequency control of standalone SEIG feeding three-phase four-wire system. IEEE Trans Sustain Energy 10(1):248–256. https://doi.org/10.1109/TSTE.2018.2832140
9. Giri AK, Arya SR, Maurya R, Chitti Babu B (2019) VCO-less PLL control-based voltage-source converter for power quality improvement in distributed generation system. IET Electr Power Appl 13(8):1114–1124. https://doi.org/10.1049/iet-epa.2018.5827
10. Sorgulu F, Dincer I (2018) A renewable source based hydrogen energy system for residential applications. Int J Hydrog Energy 43(11):5842–5851. https://doi.org/10.1016/j.ijhydene.2017.10.101
11. Sombir S, Singh M (2020) Voltage and frequency control of self excited induction generator integrated with PV system. In: IECON Proc. Industrial Electron. Conf., vol. 2020, pp. 4306–4311. https://doi.org/10.1109/IECON43393.2020.9254608.
12. Taoufik M, Abdelhamid B, Lassad S (2018) Stand-alone self-excited induction generator driven by a wind turbine. Alex Eng J 57(2):781–786. https://doi.org/10.1016/j.aej.2017.01.009
13. S. Kundu, M. Singh, and A. K. Giri, "Synchronization and control of WECS-SPV-BSS-based distributed generation system using ICCF-PLL control approach," *Electr. Power Syst. Res.*, vol. 226, no. July 2023, p. 109919, 2024, https://doi.org/10.1016/j.epsr.2023.109919.
14. Sharma R, Kewat S, Singh B (2019) Robust 3IMPL control algorithm for power management of SyRG/PV/BES-Based distributed islanded microgrid. IEEE Trans Ind Electron 66(10):7765–7777. https://doi.org/10.1109/TIE.2018.2880673
15. Sombir MS, Giri AK (2022) Implementation of in-loop filter based DSC-FLL control approach for power quality improvement in isolated microgrid. In: 2022 IEEE 31st international symposium on industrial electronics (ISIE), pp. 516–521. https://doi.org/10.1109/ISIE51582.2022.9831712.
16. Giri AKK, Arya SR, Maurya R, Babu BC (2018) Power quality improvement in stand-alone SEIG-based distributed generation system using lorentzian norm adaptive filter. IEEE Trans Ind Appl 54(5):5256–5266. https://doi.org/10.1109/TIA.2018.2812867
17. Krishnaa B, Premalatha K.A novel approach for the voltage control of self-excited induction generator - PV array assisted inverter
18. Ku TT, Li CS (2021) Implementation of battery energy storage system for an island microgrid with high PV penetration. IEEE Trans Ind Appl 57(4):3416–3424. https://doi.org/10.1109/TIA.2021.3075655
19. Kumar S, Singh B (2017) Seamless transition of three phase microgrid with load compensation capabilities. In: 2017 IEEE industry applications society annual meeting IAS 2017, vol. 2017, pp. 1–9. https://doi.org/10.1109/IAS.2017.8101757.
20. Kundu S, Singh M, Giri AK (2022) Control algorithm for coordinated operation of wind-solar microgrid standalone generation system. Energy Sour Part A Recover Util Environ Eff 44(4):10024–10044. https://doi.org/10.1080/15567036.2022.2143943
21. Kumar S, Singh B (2018) A multipurpose PV system integrated to a three-phase distribution system using an LWDF-based approach. IEEE Trans Power Electron 33(1):739–748. https://doi.org/10.1109/TPEL.2017.2665526
22. Mohamed SA, Abd M, Sattar E (2019) A comparative study of P & O and INC maximum power point tracking techniques for grid - connected PV systems. SN Appl Sci 1(2):1–13. https://doi.org/10.1007/s42452-018-0134-4
23. Ahmad R, Murtaza AF, Sher HA (2019) Power tracking techniques for efficient operation of photovoltaic array in solar applications—a review. Renew Sustain Energy Rev 101:82–102. https://doi.org/10.1016/j.rser.2018.10.015
24. Sahoo B, Routray SK, Rout PK (2021) A modified least mean square technique for harmonic elimination. In: 1st Odisha international conference on electrical power engineering, communication and computing technology ODICON 2021. https://doi.org/10.1109/ODICON50556.2021.9428970.
25. Mittal S, Singh A, Chittora P (2023) Design and development of leaky least mean fourth control algorithm for single-phase grid-connected multilevel inverter. Int J Circuit Theory Appl. https://doi.org/10.1002/cta.3746
26. Kumar PS, Chandrasena RPS, Ramu V, Srinivas GN, Babu KVSM (2020) Energy management system for small scale hybrid wind solar battery based microgrid. IEEE Access 8:8336–8345. https://doi.org/10.1109/ACCESS.2020.2964052
27. Wang Y, Tan KT, So PL (2013) Coordinated control of battery energy storage system in a microgrid. In: Asia-Pacific power and energy engineering conference APPEEC.https://doi.org/10.1109/APPEEC.2013.6837211
28. Sood K, Muthusamy E (2020) A comprehensive review on hybrid renewable energy systems. Mod Phys Lett B. https://doi.org/10.1142/S0217984920502905
29. Pathak G, Singh B, Panigrahi BK (2019) Wind-hydro microgrid and its control for rural energy system. IEEE Trans Ind Appl 55(3):3037–3045. https://doi.org/10.1109/TIA.2019.2897659
30. Kundu S, Singh M, Giri AK (2023) Adaptive control approach-based isolated microgrid system with alleviating power quality problems. Electr Power Compon Syst. https://doi.org/10.1080/15325008.2023.2239222

31. Golestan S, Guerrero JM, Vasquez JC (2017) Three-phase PLLs: a review of recent advances. IEEE Trans Power Electron 32(3):1894–1907. https://doi.org/10.1109/TPEL.2016.2565642
32. Golestan S, Guerrero JM, Vasquez JC, Abusorrah AM, Al-Turki Y (2020) All-pass-filter-based PLL systems: linear modeling, analysis, and comparative evaluation. IEEE Trans Power Electron 35(4):3558–3572. https://doi.org/10.1109/TPEL.2019.2937936
33. Qasim M, Kanjiya P, Khadkikar V (2014) Artificial-neural-network-based phase-locking scheme for active power filters. IEEE Trans Ind Electron 61(8):3857–3866. https://doi.org/10.1109/TIE.2013.2284132
34. Kalla UK, Singh B, Murthy SS (2014) Adaptive noise suppression filter based integrated voltage and frequency controller for two-winding single-phase self-excited induction generator. IET Renew Power Gener 8(8):827–837. https://doi.org/10.1049/iet-rpg.2013.0271
35. Ranjan A, Kewat S, Singh B (2021) DSOGI-PLL with in-loop filter based solar grid interfaced system for alleviating power quality problems. IEEE Trans Ind Appl 57(1):730–740. https://doi.org/10.1109/TIA.2020.3029125
36. Qiu C, Qian G, Wang S (2022) Widely linear maximum complex correntropy criterion affine projection algorithm and its performance analysis. IEEE Trans Signal Process 70:3540–3550. https://doi.org/10.1109/TSP.2022.3185903
37. Dash A et al (2022) Modified proportionate affine projection algorithm based adaptive DSTATCOM control with increased convergence speed. IEEE Access 10:43081–43092. https://doi.org/10.1109/ACCESS.2022.3169618

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

American International Journal of Research in Humanities, Arts and Social Sciences

Available online at http://www.iasir.net

ISSN (Print): 2328-3734, ISSN (Online): 2328-3696, ISSN (CD-ROM): 2328-3688

AIJRHASS is a refereed, indexed, peer-reviewed, multidisciplinary and open access journal published by International Association of Scientific Innovation and Research (IASIR), USA

(An Association Unifying the Sciences, Engineering, and Applied Research)

An Empirical Study to Enhance Leadership Qualities among Students by Improving the Social Intelligence Level of School Teachers: A Study of Select Schools in Delhi and NCR

Dr. Deepa Bisht
Associate Professor
Sivanthi College of Education,
Mehta Nagar, Kundrathur, Chennai, India.
Dr. Meha Joshi
Assistant Professor
Delhi School of Management,
Delhi Technological University, Delhi, India.

Abstract: *The framework within which the present study, research conducted was descriptive in nature and the target population in this regard is consisting of 100 schoolteachers based in urban and rural areas. Stratified Random sampling method was adopted for selecting the samples. The respondents' age varied from 25 years to 50 years. Data was collected from respondents with the help of self-administered questionnaires. For assessing the Social Intelligence level of teachers, the scale developed by Chadha and Ganeshan was used. The study highlighted that there exists a significant difference in the Social Intelligence level of teachers based on their location status. It also demonstrated that the Social Intelligence level of teachers is average. It is even posited in the study that there exists relationship among variables like Patience and Cooperativeness and therefore it can be inferred that these factors can definitely help predict the Social Intelligence level better.*

Keywords: *Social Intelligence of Schoolteachers, Social Intelligence and leadership quality.*

I. Introduction

In the recent past, social intelligence has become a much talked about term. It is with the advent of leadership issues gaining momentum that it has been realised that Social Intelligence related concepts assume importance. The present study relates to social intelligence in context of school teachers. It is well established fact that students see their faculty as their role model. There are many lessons of life and character building which are primarily learnt from teachers. From this it is evident that if efforts are put in, to raise the level of social intelligence among teachers, the tone for social intelligence among the students may automatically be set.

II. Review of Literature

In 1998 Danial Goleman came up with the exploration of the idea of the Emotional Intelligence (EI) and leadership in his book "What makes a leader?" Since then, vital role of the components of EI; like empathy and self-knowledge has been a subject of talk among diverse communities. But in recent past, researches in the field of 'social- neuroscience: the study of what happens in brain when people interact', has unveiled much of truth about 'what makes a good leader'; and the concept of Social Intelligence (SI) comes into limelight. Though, SI has its origin in Thorndike's (1920) division of intelligence into three facets: the ability to understand and manage ideas (abstract intelligence), concrete objects (mechanical intelligence), and people (social intelligence). The term SI, referred to the person's ability to understand and manage other people and to engage in adaptive social interaction (Thorndike, 1920). More recently, however, Cantor and Kihlstrom (1987) redefined SI as "the individual's fund of knowledge about the social world". A more relationship-based construct for assessing leadership is social intelligence, which we define as a set of interpersonal competencies built on specific neural circuits (Goleman & Boyatzis, 2008). Various researches on social intelligence indicate that how it is linked with everything from student achievement, workplace success to general wellbeing (Bar-on, 2005; Druskatet et.al. 2006). Investigation on various aspects of social intelligence have shown that a high level of social intelligence contributes towards the person's ability to meet the demands of everyday functioning as well as for shouldering responsibility of the welfare of self and others. According to Cantor & Kihlstrom (1987), social intelligences specifically geared for solving the problems of social life, and in particularly managing the life tasks, current concerns or personal projects which the person selects for him or herself, or which other people impose on him or her from outside.

Many researchers reported that social intelligence is multidimensional and distinguished from general intelligence domains (Wong, Day, Maxwell, & Meara, 1995; Jones & Day, 1997; Weis & Süb, 2007) which contributes towards effective leadership. Zirkel (2000) believed that social intelligence is closely related to one's own personality and individual behaviour. SI is the part of individual's mental ability which generates the capacity to adapt him to the society, as such relations are essential for his existence. It is human capacity to understand what is happening in the world and respond to that understanding in personal and social level in effective manner. SI has a great importance for the teachers to interact with students effectively, for better understanding of their personality so as they can be harnessed for shaping up a better personality makeup. Balaghat and Pour (2014) investigated the relationship between social intelligence of high school managers of district-2 of Zahedan and their leadership style and found a positive and significant relationship between SI and relation-oriented managers. Albrecht (2006) suggested that social intelligence is a prerequisite for teachers; education system and teachers respect the rules and behaviours associated with high social intelligence. Jeloudar and Goodarzi (2012) found a significant relationship between teachers' social intelligence and their job satisfaction; and, teachers' social intelligence and five factors of job satisfaction: nature of the work itself, attitudes towards supervisors, relations with co-workers, opportunities for promotion, work condition in the present environment, but the relationship with one factor (salary and benefit) of job satisfaction is low and negligible. Sarvamangala (2012) observed that socio-demographic factors such as gender, age, caste and marital status etc. do not have any influence on social intelligence of school teachers. Toor (2013) found significant difference in social intelligence of male and female secondary school teachers of punjab. McQuade (2013) demonstrated that principals and teachers view social intelligence critical to educational leadership and teachers' improvement. Renugadevi and Nalinilatha (2014) found no significant relationship between social intelligence and teaching competency. Mahaboobvali and Vardhini (2016) studied the social intelligence of school teachers with respect to gender and age; and found no gender influence on social intelligence of teachers whereas a significant variation was found among the age groups of the teachers towards the social intelligence.

Based on the premise that leadership traits can be acquired by raising the level of social intelligence among the students and teachers, the research has been conducted with the following objectives.

1. To find out the social intelligence level of school teachers.
2. To find out the relationship among the various factors contributing to the social intelligence level viz. Patience, Confidence, Cooperativeness, Sensitivity, Recognition in the social environment, Tactfulness, Sense of humour and Memory.
3. To test that whether there is any significant difference in the Social intelligence level of teacher based on the urban or rural areas where schools are located.

III. Hypotheses of the Study:

H1: The School teachers possess high social intelligence.
H2: There exists a positive relationship among the various factors pertaining to social intelligence level.
H3: There is no significant difference in the social intelligence level of teachers based on their locality.

IV. Research Design:

The framework within which the research was conducted was basically descriptive in nature and the target population in this regard is defined as consisting of 100 school teachers based in urban and rural areas. The research was conducted on school teachers. Stratified Random sampling method was adopted for selecting the samples. The respondents' age varied from 25 years to 50 years. Survey methods was used to collect data from respondents with the help of self-administered questionnaires.

A scale comprising of items clubbed into 8 factors was used to determine the Social intelligence level of teachers on 3 point Likert scale. In the process 8 factors were considered for determining the Social intelligence level of teachers. Social intelligence inventory by Chadha and Ganesan were used . The process of data collection took a period of over 4 months. Participation of teachers in the survey was voluntary. Socail intelligence was measured using item scales comprising 5 parts with total of 66 questions. For assessing the Social Intelligence level of teachers, the scale developed by Chadha and Ganesan was used. This was chosen as it is considered as a reliable inventory for assessing the Social intelligence level. On the inventories, reliability analysis was done by using Cronbach alpha test and the score came out to be more than .7 (.85in particular) which indicated that the statements are internally consistent. One way ANOVA was used to measure the significant difference between the factors based on the location status of schools. Karl Pearson's Correlation was used to find the relationship of various factors of social intelligence among each other and Total Social Intelligence score.

For six dimensions of Social Intelligence viz. Patience, Confidence, Sensitivity, Cooperativeness, sense of humour and Recognition of social environment, score of 1 indicated lack of confidence. In case of a dimension tactfulness, the response was in the form of 'yes ' or 'no'. A high score indicated high social intelligence.

The scale comprised 66 items . Each item had options namely “Very high ”, ”high”, ”Average”, “Low”, ”very low”. An individual score is the sum of all the scores of 66 items. The score ranged from 66 to 138. As per manual, the level of social Intelligence has been defined as below:

High Social Intelligence 113 and above
Average Social Intelligence 57 upto 112
Low Social Intelligence 36 to 56

V. Data Analysis

Table 1: Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Social_Int	100	87	123	109.03	7.780
Valid N (listwise)	100				

The data was collected from 100 school teachers and their mean was computed. The mean score of Social Intelligence was computed and the results showed that the school teachers had average social intelligence (According to the interpretation of the scores of the inventory, anyone having a score between 57 to 112 is said to possess Average Social Intelligence.

In the inventory, Social Intelligence is computed as a measure of additive scores of the factors viz. Patience, Confidence, Cooperativeness, Sensitivity, Recognition in the social environment, Tactfulness, Sense of humour and Memory.Therefore the hypothesis stating that social intelligence level of school teachers is high is rejected.

The details of the scores obtained by teachers on individual factors contributing to Social Intelligence level have been given as in Table 2.

Table 2: Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Total_ Social_ Int	100	87	123	109.03	7.780
Patience	100	11	26	20.71	2.931
Confidence	100	14	30	19.90	2.676
Cooperativeness	100	22	31	27.27	2.403
Sensitivity	100	15	25	20.76	2.775
Recog_Soc_env	100	0	3	1.43	.795
Tactfulness	100	1	6	4.02	1.310
Sense_of_humour	100	1	8	4.72	1.676
Memory	100	6	13	10.15	1.828
Valid N (listwise)	100				

Table 3: Correlation between various factors of Social Intelligence

Correlations										
		Patience	Cooperativeness	Confidence	Sensitivity	Recog_Soc_env	Tactfulness	Sense_of_humour	Memory	Total_Social_Int
Patience	Pearson Correlation	**1**	.206*	.251*	.202*	-.041	.133	-.072	.025	.610**
	Sig. (2-tailed)		.040	.012	.043	.683	.187	.475	.804	.000
Cooperativeness	Pearson Correlation	.206*	**1**	.149	.111	-.257**	.338**	.139	.191	.611**
	Sig. (2-tailed)	.040		.140	.270	.010	.001	.167	.057	.000
Confidence	Pearson Correlation	.251*	.149	**1**	.232*	.020	-.057	-.137	-.168	.492**
	Sig. (2-tailed)	.012	.140		.020	.840	.573	.174	.094	.000
Sensitivity	Pearson Correlation	.202*	.111	.232*	**1**	.015	-.021	-.058	.220*	.569**
	Sig. (2-tailed)	.043	.270	.020		.881	.837	.566	.028	.000
Recog_Soc_env	Pearson Correlation	-.041	-.257**	.020	.015	**1**	-.086	.129	-.017	.026
	Sig. (2-tailed)	.683	.010	.840	.881		.395	.200	.866	.800

Tactfulness	Pearson Correlation	.133	.338**	-.057	-.021	-.086	**1**	-.108	.121	.340**
	Sig. (2-tailed)	.187	.001	.573	.837	.395		.286	.230	.001
Sense_of_humour	Pearson Correlation	-.072	.139	-.137	-.058	.129	-.108	**1**	-.105	.101
	Sig. (2-tailed)	.475	.167	.174	.566	.200	.286		.299	.316
Memory	Pearson Correlation	.025	.191	-.168	.220*	-.017	.121	-.105	**1**	.339**
	Sig. (2-tailed)	.804	.057	.094	.028	.866	.230	.299		.001
Total_Socil_Int	Pearson Correlation	.610**	.611**	.492**	.569**	.026	.340**	.101	.339**	**1**
	Sig. (2-tailed)	.000	.000	.000	.000	.800	.001	.316	.001	

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

Table 3 shows Karl Pearson's coefficient of correlation among the various factors of Social intelligence viz. . Patience, Confidence, Cooperativeness, Sensitivity, Recognition in the social environment, Tactfulness, Sense of humour and Memory. The correlation scores obtained by different factors of social intelligence and social intelligence are as given in the Table 3.

The Table clearly shows that patience(r=.610) and cooperativeness (r=.611)have strong positive correlation with social intelligence among the others. It can be seen that Sensitivity, Tactfulness and confidence also show significant positive correlation with spiritual intelligence level but their magnitude of relationship is fairly lesser than the Patience and Cooperativeness. The factors showing positive relationship with Social intelligence were tested at .01 level of significance. Sense of humour and Memory were found to be insignificant to determine Social Intelligence level of the teachers. Therefore, it can be inferred that the Patience, Cooperativeness, Confidence and sensitivity can best predict the Social Intelligence of the teachers.

Table 4
Analysis of Variance for Various Factors of Social Intelligence With Respect to Location of the Schools (Remote and Urban)

Factor Name	Remote		Urban		F value	P value
	Mean	SD	Mean	SD		
Patience	4.31	.58	3.29	.57	.56	.007*
cooperativeness	4.74	.697	3.98	.83	.17	.004*
confidence	4.19	.56	4.18	.72	.25	.77
Sensitivity	4.20	.628	4.27	.58	.77	.46
Recognition in the social environment	2.52	.57	4.79	.60	4.72	.03**
Sense of humor	4.24	.70	4.15	.73	1.07	.34
Memory	4.35	.63	4.24	.84	4.48	.98

*significant at.01 level
**significant at .05 level

One way ANOVA was performed to find that whether there is any significant difference in the social intelligence level among teacher with respect to the location at which they are placed. The result showed that there is a significant difference at 99% confidence level in the Social Intelligence level based on location status of the schools. More specifically the intelligence varied for the factors viz. Patience and Cooperativeness and at 95% confidence level, it is established that recognition to social environment also varies with the location status. The null hypothesis stating that there is a significant difference based on the location status is rejected

V. Findings and Recommendations:

The study highlighted that there exists a significant difference in the Social Intelligence level of teachers based on their location status. It also demonstrated that the Social Intelligence level of teachers is average. It is even posited in the study that there exists relationship among variables like Patience and Cooperativeness and therefore it can be inferred that these factors can definitely help predict the Social Intelligence level better. It is recommended that schools should try to induce Social Intelligence enhancing modules in the course curriculum for teachers and such workshops should become an integral part of the faculty development programmes.

References:

[1]. Albrecht, K.(2006). *Social Intelligence: The New Science of Success.* San Francisco, CA: Jossey-Bass.
[2]. Balaghat, S. R. and Pour, L.A. (2014). Investigation of the relationship between social intelligence and leadership style of high school managers at dictrict-2 in Zahedan. European Online Journal of Natural and Social Sciences, Vol.3(1), p.p. 140-149.
[3]. Bar-On, R. (2005). The Bar-On model of emotional-social intelligence. In P. Fernandez-Berrocal & N. Extremer (Eds.), Special Issue on Emotional Intelligence. *Psicothema*
[4]. Cantor, N., & Kihlstrom, J. F. (1987). *Personality and social intelligence.* Englewood Cliffs, N.J: Prentice-Hall.
[5]. Druskat, V., Sala, F., & Mount, G. (2006). *Linking emotional intelligence and performance at work.* Mahwah, NJ: Lawrence Erlbaum Associates.
[6]. Goleman, D. & Boyatzis, R. E. (2008). Social intelligence and biology of leadership. Harverd Business Reivew. Retrived from https://hbr.org/2008/09/social-intelligence-and-the-biology-of-leadership
[7]. Jeloudar, S.Y. and Goodarzi, F.L. (2012). The Relationship between Social Intelligence and Job Satisfaction among MA and BA Teachers. International Journal of Education and Science 4(3): 209-213.
[8]. Jones, K., & Day, J. D. Journal of Educational Psychology, 1997, 89(3), Pp: 486-497.
[9]. Mahaboobvali, K and Vardhini, S. V. (2016). Social Intelligence of Secondary School Teachers with Respect to Their Gender and Age. The International Journal of Indian Psychology, vol 3, issue 2, no 2.

Analysing Customer Satisfaction Towards Personal Loans: Evidence from Banking Industry

Divya[1], Deepti Aggrawal[2] and Adarsh Anand[1,*]

[1]*Department of Operational Research, University of Delhi, Delhi 110007, India*
[2]*USME, Delhi Technological University, Delhi 110095, India*
E-mail: divya@or.du.ac.in; deeptiaggrawal@dtu.ac.in; adarsh.anand86@gmail.com
[*]*Corresponding Author*

Received 17 April 2024; Accepted 21 May 2024

Abstract

With an emphasis on banking industry, this study intends to examine customer satisfaction levels of those who have taken out personal loans from banking institutions. Customer satisfaction is a crucial indicator of success and longevity in the banking industry. This study explores the specifics of customer satisfaction among recipients of personal loans; wherein, the primary goal is to use operational research methods to analyse the intricate topography of customer satisfaction in the personal loan sector. The study's secondary objective is to demonstrate how factor analysis, a well-known dimensionality reduction method, may combine these disparate traits into a condensed set of essential components. The work presents practical insights to guide strategic choices intended to improve customer satisfaction and foster long-term growth in the banking industry, focusing on Banking sector.

Keywords: Banking sector, customer satisfaction, factor analysis, marketing research, personal loan.

Journal of Graphic Era University, Vol. 12_1, 1–22.
doi: 10.13052/jgeu0975-1416.12110
© 2024 River Publishers

1 Introduction

The banking sector provides individuals and businesses with vital financial services, making it a pillar of contemporary economies. Personal loans are critical to these services, providing customers access to money for various uses, including debt reduction, home improvement, education, and medical costs [1]. Understanding and evaluating consumer satisfaction with these financial products is crucial for banks to stay competitive and sustain long-term relationships with their clients as the demand for personal loans rises [2]. Customer satisfaction is a crucial indicator of how successfully a bank meets or exceeds its clients' requirements and expectations. Customer satisfaction in the context of personal loans includes several factors, such as the simplicity of the application process, interest rates, loan terms, level of customer service, openness, and general borrowing experience. The efficiency of banking services can be better understood by looking at customer satisfaction with personal loans. This will enable institutions to customize their products to suit the needs of their clientele better and increase retention rates. In the context of company strategy and consumer-centric methodologies, there is a close relationship between marketing research and customer satisfaction [3].

Customer satisfaction is the result of providing goods and services that meet or exceed the consumer's desires and expectations [4]. Customer satisfaction is a concept in the marketing research field, so it is crucial to understand what marketing research is. Marketing research is essential for comprehending customer preferences, behaviours, and perceptions [5]. Marketing research specifies the information required to address many issues, like designing the method for collecting information, managing and implementing the data collection process, analysing the results, and communicating the findings and their implications [6]. It systematically gathers, records, and analyses data about marketing products and service issues. Marketing research aims to identify and assess how changing elements of the marketing mix impact customer behaviour [7]. The term is commonly interchanged with market research; however, expert practitioners may wish to distinguish between market research is concerned specifically with markets while marketing research is concerned specifically about marketing processes. Also, this process is often partitioned into two sets of categorical pairs: target market, Consumer marketing research, and Business-to-business (B2B) marketing research, or by methodological approach, Qualitative marketing research, and Quantitative marketing research [8]. Consumer marketing research is a form of applied sociology that concentrates on understanding

consumers' preferences, attitudes, and behaviours in a market-based economy. It aims to understand the effects and comparative success of marketing campaigns. Arthur Nielsen pioneered consumer marketing research as a statistical science with the founding of the AC Nielsen Company in 1923 [9].

1.1 Concept of Customer Satisfaction

A fundamental idea in marketing and business management is that customer satisfaction is the level of fulfilment or contentment that consumers feel after interacting with a business, service, or product [10]. It includes their thoughts, emotions, and assessments about several facets of their experience, such as the Caliber of the goods, the promptness of the service, and their general interaction with the business.

Important facets of the customer satisfaction notion consist of:

- Expectations vs. Experience: The degree to which a customer's experience meets or surpasses their initial expectations is a common way to measure customer satisfaction. When a product or service exceeds clients' expectations, they will probably be pleased. If, on the other hand, the experience doesn't live up to expectations, disappointment could result. Customers' perceptions of quality are influenced by their demands, experiences, and preferences [11]. Subjective elements like dependability, effectiveness, beauty, and value for money are important in determining quality. When goods or services match their expectations for quality, customers are most likely to be happy.
- Service Delivery: The way goods or services are delivered can greatly impact customer satisfaction. Prompt and courteous service, customized attention, efficient problem resolution, and general attentiveness influence positive customer experiences and better satisfaction levels. Experiences with the brand at different touchpoints during the consumer journey shape impressions of it [12].
- Emotional Engagement: Satisfaction with customers results from both rational and emotional factors. Positive feelings like joy, delight, and trust foster stronger emotional ties between consumers and brands and are linked to high satisfaction levels [13]. Engaging customers emotionally increases their advocacy and loyalty, encouraging repeat business and positive word-of-mouth referrals.
- Loyalty and Retention: Happy consumers are more likely to stick with a brand over time and make repeat purchases of its goods or services [14]. Additionally, they are more likely to tell others about the company,

which helps with customer acquisition and retention initiatives. Consequently, establishing and sustaining client happiness is essential to long-term company success.

- Continuous Improvement: Due to shifting consumer tastes, industry developments, and competitive pressures, customer satisfaction is dynamic and prone to change. Companies must use performance indicators, market research, and feedback channels to monitor and evaluate customer satisfaction. Businesses may improve customer satisfaction and keep a competitive edge by recognizing areas for improvement and making necessary adjustments to meet changing needs [15]. Customer satisfaction measures how well a client's wants and expectations are satisfied when contacting a business, service, or product.

The banking industry's primary goal is to offer financial services and solutions that satisfy customers' requirements and expectations, and customer satisfaction is intrinsically linked to this goal [16].

1.2 Banking Sector

In the banking sector, many factors contribute to customer satisfaction: service quality, product selection, ease of use, accessibility, and overall customer experience. Maintaining long-term connections with their client and thriving in today's competitive landscape need banks to understand and improve customer satisfaction [17]. The calibre of services offered is one of the main factors influencing client satisfaction in the banking industry. This includes how quickly bank personnel respond to you, how smoothly transactions go, how accurate the information is, and how professionally all of the bank's employees act. Banks must meet various consumer needs, from simple savings and checking accounts to more complicated financial products, including loans, mortgages, and investment possibilities. Customers' satisfaction levels can be raised by offering creative and personalized products at transparent and upfront competitive prices. In the banking industry, accessibility and convenience are also significant determinants of client happiness [18]. Customers now demand smooth online and mobile banking experiences with features like simple account access, bill payment, fund transfers, and round-the-clock customer service, thanks to the development of digital banking technologies. User-friendly interfaces, ATM networks, and branch availability all add to the convenience and happiness of customers. Additionally, relationship management and tailored communication can significantly impact client satisfaction in the banking industry. Proactive relationship managers, tailored advice, and

attentive service can foster customers' trust and loyalty, ultimately increasing their satisfaction with the bank [19].

Customer satisfaction is a complex topic in the banking industry that includes a range of factors such as product choices, ease, trust, and customized relationships. Banks are better positioned to build enduring connections, encourage client loyalty, and prosper in a market that is becoming increasingly competitive when they prioritize comprehending and meeting their customers' needs and preferences. To analyse customer satisfaction with personal loans, the author used factor analysis as a tool. Factor analysis offers a strong foundation for comprehending the fundamental aspects of client contentment and locating workable solutions for raising satisfaction levels inside and outside the banking industry. Factor analysis identifies discrete client groups with comparable satisfaction patterns, which banks can use to segment their customer base better. By customizing their product offers and marketing methods to the distinct requirements and tastes of various client segments, banks can increase customer retention and loyalty through segmentation. In conclusion, factor analysis and marketing research are essential elements of strategic decision-making in the banking industry. Banks can gain valuable insights into the factors that drive customer satisfaction, optimize their offerings, and improve overall customer experiences by utilizing factor analysis techniques and marketing research methodologies. This can lead to developing long-term relationships and a competitive advantage in the market. The study employs OR methods and factor analysis to obtain practical insights to guide strategic choices intended to improve customer satisfaction and foster long-term growth in the banking industry, focusing on Yes Bank.

The rest of the study has been systematized as follows: Section 2 gives an extensive literature review on customer satisfaction in the banking sector. Sections 3 and 3.1 describe the Research Methodology of this study. Section 3.2 interprets the research methodology used, which is Factor Analysis. The final results of this study are described in Section 4. The author has concluded this study in Section 5, and future scopes and limitations of the study are also discussed in the Sections 5.1 and 5.2.

2 Literature Review

This section has extensively reviewed the literature on customer satisfaction in the banking sector to understand the comprehensive construction and research objective. The banking sector mainly relies on customer service

engagement and perspective, and customers, the backbone of the economy, have a significant role in determining their satisfaction.

Parasuraman et al. [20] presented a multi-item scale technique for measuring customer perception in the service industry in their study. They have emphasized the importance of studying customer perception and satisfaction. Hanif et al. [21] studied various factors that affect customer satisfaction. Singh et al. [22] empirically studied customer satisfaction and universal banks. The customer perception and satisfaction towards home loans have been analysed by Chaudhary and Janjhua [23]. Dawar [24] identified critical factors influencing customer satisfaction in banking using factor analysis. Out of the initial 22 attributes, factor analysis was reduced to 6 factors. Linking the banking industry with loan procedure variables, Jain et al. [25] studied the impact of demographic characteristics on customer satisfaction. Customer relationship management is based on increasing customer satisfaction analysed by Tao [26]. In the same year, Kapur et al. [27] presented a study for evaluation based on the expectations of customers and their satisfaction. Using logistic regression in 2016, Anand and Bansal predicted customer satisfaction and dissatisfaction. Begum et al. [28] used a confirmatory factor analysis to determine customer satisfaction with corporate loans. Later, Anand et al. [29] quantified the adopters as satisfied users and dissatisfied customers. In the year 2021, Tien et al. [30] studied customer satisfaction at a bank and used factor analysis for dimension reduction. Agag et al. [31] studied marketing analytics to understand the relationship between customer agility and customer satisfaction. These studies offer a solid framework for examining consumer satisfaction with personal loans in the banking industry. This aligns with the study's goals of employing operational research (OR) to evaluate customer satisfaction with a specific focus on Yes Bank.

3 Research Methodology

For this study, the authors conducted a survey on customers of Yes Bank at Hudson Lane, Kingsway Camp, New Delhi-110009, India, in the age group of 21–60 years, to analyse how satisfied they are with the services of the bank under consideration. They were intentionally chosen as this market is vital, easily influenced by its surroundings, and good at buying power. Two hundred eleven customers were administered randomly, and the area where questionnaires were distributed was the University of Delhi.

Table 1 is a representation of the customers as per their educational/ qualification level:

Table 1 Frequency table of customers' qualification

Education Level	Count of Education Level
Doctorate	6
High School	4
Other	6
Postgraduate	43
Secondary School	5
Undergraduate	146
Grand Total	**210**

In this study, customers were asked to rate their satisfaction with various attributes connected to the dissemination of personal loans. Apart from the different demographic questions, they rated various attributes from 1 to 5, with one being the least satisfied and 5 being the highest satisfaction with a particular attribute.

The authors calculated the weighted mean of those attributes to determine which attributes have the most impact on making the brand more qualitative and valuable among customers. The table below showcases the number of persons catered in the different categories based on their occupation.

Table 2 shows that most of the respondents are private sector employees, followed by the government sector, and so on.

Table 3 describes the various reasons why particular loans have been taken from the bank.

Here, it can be seen that most of the respondents took a loan to finance a new venture after that for medical emergencies, debt consolidation, and so on.

After fetching useful information from the demographics, the next step is moving from demography toward dimensionality reduction. As described

Table 2 Occupation-wise division of the customers

Occupation	Count of Education Level
Business	33
Government Employee	62
Home Maker	4
Other	23
Private Job	70
Retired	1
Self Employed	18
Grand Total	**211**

Table 3 Reasons for availing of personal loans

Purpose for Availing of Personal Loan	Count of Education Level
Debt consolidation	35
Finance a new venture	52
Home remodelling	33
Medical Emergencies	38
Other	25
Vacation Costs	6
Wedding Expenses	22
Grand Total	**211**

above, the various attributes under consideration are presented to the customers, and they are asked to rate them based on their satisfaction level. Before moving toward the analysis, it is essential to understand factor analysis. In the next section, the author briefly describes the methodology used – Factor Analysis.

3.1 Factor Analysis

A statistical method called factor analysis is used to find underlying dimensions or factors that account for patterns of correlations between a group of observed data. Factor analysis can be used in customer satisfaction to determine the main aspects influencing overall satisfaction and the underlying structure of variables connected to satisfaction.

The following actions can be conducted to perform factor analysis on customer satisfaction data:

- Step 1: Data collection: Compile information on a range of customer satisfaction factors, including product characteristics, the calibre of the services, cost, ease of use, and overall experience. Feedback forms, interviews, and surveys can all be used to get this information.
- Step 2: Choose the variables (or survey questions) that best capture the many facets of customer satisfaction. These variables should encompass a broad range of satisfaction parameters and be pertinent to the area of interest.
- Step 3: Data Preparation: Ensure the data satisfies the requirements of factor analysis, including continuous variables, a sizable enough sample size, and no problems with multicollinearity. It can be necessary to attribute or exclude missing data.

- Step 4: Factor Extraction: Use methods for factor extraction, such as joint factor analysis or principal component analysis (PCA), to find the underlying factors that account for the patterns of correlations between the variables. The aim is to reduce the number of dimensions in the data while maintaining the maximum amount of information.
- Step 5: Factor Rotation: After extracting the factors, use rotation techniques like varimax or rotation to create a more straightforward and understandable factor structure. Rotation makes it simpler to see how variables and factors relate to one another.
- Step 6: Factor Interpretation: To understand the results, look at the factor loadings, which show how strongly and in which direction each variable is related to the underlying factors. High-loading variables on a given factor are thought to be closely related to that factor.
- Step 7: Naming & Labelling: Based on the analysis of factor loadings, give the extracted factors that represent the fundamental aspects of customer satisfaction appropriate names or labels. These labels should reflect each component's variables' general themes or traits.
- Step 8: Assessment of Validity and Reliability: Determine the extracted variables' internal consistency and construct validity to assess their validity and reliability. The reliability of the factors can be evaluated using methods like Cronbach's alpha, and the model fit and validity of the factor structure can be examined using confirmatory factor analysis (CFA).
- Step 9: Application and Interpretation: Apply the final factor structure to customer satisfaction and use the insights gained to improve customer experiences. The factors identified have the potential to impact strategic choices about product creation, marketing tactics, improving customer service, and general business operations.

The authors have used 16 variables to perform a factor analysis to analyse customer satisfaction with personal loans. Table 4 represents the descriptive statistics performed on these 16 variables.

But before proceeding with factor analysis, we need to test the KMO and Bartlett's tests to determine the data's suitability. This test measures sampling adequacy for each variable in the model and for the complete model. The statistic measures the proportion of variance among variables that might be common variance. The lower the proportion, the more suitable the collected data is for factor analysis.

Table 4 Descriptive statistics

	Mean	Std. Deviation	Analysis N
Debt to Income Ratio	3.50	1.319	208
Top Up Loan	3.74	1.155	208
Employment History	3.39	1.506	208
Efficient Staff	3.30	1.040	208
Monthly installment	3.41	1.201	208
Rate of Interest	4.09	1.147	208
Time taken for the verification	3.13	1.062	208
Credit Score (CIBIL Score)	3.72	1.448	208
E-Verification	3.53	0.962	208
Penalty fees	2.99	1.026	208
Loan Repayment	2.89	1.331	208
Following RBI Guidelines	3.23	0.994	208
Facility of loan extension	3.77	1.033	208
Quick Loan Approval facility	3.54	1.289	208
Documentation Process	3.08	1.166	208
Low File Closure Charge	3.23	1.356	208

Table 5 KMO and Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.825
Bartlett's Test of Sphericity	Approx. Chi-Square	1035.129
	df	120
	Sig.	$<$0.001

As a rule of thumb for interpreting the statistics:

- KMO values between 0.8 and 1 indicate the sampling is adequate.
- KMO values less than 0.6 indicate inadequate sampling, and that remedial action should be taken. Some authors put this value at 0.5, so use your judgment for values between 0.5 and 0.6.
- KMO values close to zero mean there are large partial correlations compared to the sum of correlations.

For the collected data, the KMO value is 0.825, which can be seen in Table 5, which falls into the excellent range, so we can be confident that the sample size is adequate for factor analysis. Bartlett's test is a statistic used to examine the hypothesis that the variables are uncorrelated in the population. From the data, we have the KMO value of $0.825 > 0.5$. Thus, factor analysis is appropriate for this case. Hence, we can further proceed with the analysis.

Table 6 Communalities

	Initial	Extraction
Debt to Income Ratio	1.000	0.584
Top Up Loan	1.000	0.472
Employment History	1.000	0.607
Efficient Staff	1.000	0.694
Monthly installment	1.000	0.644
Rate of Interest	1.000	0.762
Time taken for the verification	1.000	0.504
Credit Score (CIBIL Score)	1.000	0.419
E-Verification	1.000	0.679
Penalty fees	1.000	0.684
Loan Repayment	1.000	0.584
Following RBI Guidelines	1.000	0.637
Facility of loan extension	1.000	0.537
Quick Loan Approval facility	1.000	0.575
Documentation Process	1.000	0.399
Low File Closure Charge	1.000	0.316

The foundation of factor analysis is to study the various variables under a limited set of technical definitions, which can then be presented to management. Around 16 different variables were considered, and the relevant queries regarding their satisfaction level were sent to the customers. Some findings are presented below in Table 6.

The amount of variance in each variable that is accounted for is shown in this table. Estimates of the variance in each variable accounted for by all components or factors are known as initial communalities. Estimates of the variance in each variable that the components account for are known as extraction communalities. The factors explain all variance, and there is only one communality prior to extraction because there are precisely as many factors as there are variables. Unfortunately, some of the components are eliminated following extraction, which results in the loss of some information. While they can only account for a portion of the variance in the data, the retained components cannot fully explain all of it. Following that, four elements are removed and kept (as shown in Table 7); the communality in the above table is 0.58 for the Debt-to-income ratio, 0.47 for Top-Up loan, and so on. Four factors have been retrieved from the 16 variables; factors with eigenvalues greater than one are presumed to be extracted. The above

Table 7 Total variance explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	6.402	40.011	40.011	6.402	40.011	40.011	5.709
2	1.283	8.019	48.030	1.283	8.019	48.030	1.334
3	1.222	7.637	55.667	1.222	7.637	55.667	4.289
4	1.012	6.328	61.995	1.012	6.328	61.995	1.222
5	0.816	5.097	67.093				
6	0.776	4.851	71.944				
7	0.713	4.459	76.402				
8	0.694	4.339	80.741				
9	0.561	3.509	84.250				
10	0.545	3.403	87.653				
11	0.443	2.766	90.419				
12	0.397	2.479	92.899				
13	0.375	2.347	95.245				
14	0.312	1.953	97.198				
15	0.255	1.596	98.794				
16	0.193	1.206	100.000				

table (all values are called communality in the above table) shows that the communality is 40% for variable 1, 48% for variable 2, and so on after the factors are extracted and retained. Finding the extracted variables, their Eigenvalues, and the cumulative percentage of variance is the first step in evaluating the results. The cumulative percentage column shows that the four factors extracted—information from the 16 original variables—account for 55% of the variance overall.

Table 8 represents the Rotated Component Matrix, and Table 9 represents the Component Matrix. These tables are required to interpret or summarize the results. These tables report the factor loadings for each variable on the factors or components after rotation. Each number represents the partial correlation between the item and the rotated factor.

The authors are able to reduce the 16 variables to 4 factors that are free from redundancy. After completing the statistical computations factoring, the next step is interpreting these factors. This is achieved by inspecting the pattern of high and low loading of each of the factors on the variables.

Table 8 Rotated component matrix

	Component 1	Component 2	Component 3	Component 4
Debt to Income Ratio			0.720	
Top Up Loan		0.590		
Employment History			0.745	
Efficient Staff	0.773			
Monthly installment				0.557
Rate of Interest		0.786		
Time taken for the verification	0.598			
Credit Score CIBIL Score			0.585	
E- Verification	0.752			
Penalty fees	0.810			
Loan Repayment				0.752
Follow RBI Guidelines	0.731			
Facility of loan extension		0.718		
Quick Loan Approval	0.601			
Documentation Process				
Low Foreclosure Charge	0.537			

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.[a]

a. Rotation converged in 7 iterations.

Table 9 Component transformation matrix

Component	1	2	3	4
1	0.816	0.345	0.442	0.140
2	−0.381	0.915	−0.050	0.122
3	−0.244	−0.202	0.325	0.891
4	−0.360	−0.049	0.835	−0.414

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

3.2 Interpretation of Factor Analysis

The results of those above require some discussion. From Table 7 of the total variance explained, it can be seen that the total 16 attributes can be clubbed under the four factors in totality. They can be given specific names to aid the management so that instead of studying the 16 attributes separately, only

four factors can be worked upon. Based on the values obtained in the rotated component matrix, around eight attributes can be clubbed under one canopy based on the values of the factor loadings.

Factor 1: Loan processing efficiency

Attributes	Factor loading
Efficient staff	0.773
Time taken for verification	0.598
E-Verification	0.752
Penalty fees	0.810
Follow RBI Guidelines	0.731
Quick loan Approval	0.601
Low foreclosure charge	0.537

Factor 1 mainly comprises the personal loan processing efficiency attributes influencing the respondents to take a loan. All the attributes have high factor loading. The attribute 'Penalty fees' has the highest factor loading of 0.810, indicating that this factor influences this factor the most.

Similarly, other factors can be created and named accordingly.

Factor 2: Revolving Credit

Attributes	Factor loading
Top-Up Loan	0.590
Rate of Interest	0.786
Facility of loan extension	0.718

Factor 2 mainly comprises the Revolving Credit attributes of the personal loan that influence the respondents to purchase. All the attributes have high factor loading. The attribute 'Rate of Interest' has the highest factor loading of 0.78, which indicates that this attribute influences this factor the most.

Factor 3: Financial Liabilities and Employment History

Attributes	Factor Loading
Debt to income ratio	0.720
Employment history	0.745
Credit Score	0.585

Factor 3 comprises mainly the Financial and employment attributes of the personal loan that influence the respondents' purchase. All the attributes have higher factor loading values. The attribute 'Employment History' has the highest factor loading of 0.74, which indicates that this attribute influences this factor the most. All the attributes within this factor have positive loading.

Factor 4: Payment Mode

Attributes	Factor loading
Monthly installment	0.557
Loan Repayments	0.752

Factor 4 mainly comprises the Payment attributes of the personal loan that influence the respondents to purchase. All the attributes have high factor loading. The attribute 'Loan Repayments' has the highest factor loading of 0.752, indicating that this factor influences this factor the most.

4 Result Discussion

The profile of the related respondents was recorded, wherein it can be seen that out of the 211 customers interviewed for this survey, 75.71% were found to be male, and 25.29% were found to be female. It can be seen that the majority of the loan wishers here are male as loan requirements for business purposes to finance a new venture are more demanded by male people and for medical emergencies, too. To analyze the age buckets under which the customers that were interviewed were falling, the age profiling of the respondents was done. It came out that around 59% of the respondents were falling under the age criteria of 31-40 years, making it the majority, whereas 29% were under 21–30 years, and the rest were under 8.53% and 2.37%, respectively. This shows that most of the customers have proper knowledge about the loan scheme they will take. An educated person makes wiser choices and is aware of the market scenarios. Here, the education level of the respondents showed the results depicting the level of qualification that respondents had. It was seen that 69% of the respondents were graduated and fell under the majority section, with 20% of the customers being postgraduate. This shows that most customers are educated and know about market scenarios well.

4.1 Implications for Practice

The study's conclusions about the banking sector's customer satisfaction with personal loans have important practical implications. First off, banks may better serve their customers by customizing their services by knowing the important factors that influence customer happiness, such as interest rates, loan conditions, and the calibre of customer care. To improve customer delivery, this can entail updating loan products, expediting the application process,

or funding employee training. Second, pinpointing areas of low customer satisfaction gives banks useful information for enhancing their operations. Banks may proactively improve their service offerings and cultivate customer happiness and loyalty by using feedback tools to gain a deeper understanding of client preferences or addressing frequent pain points encountered during the loan application process.

5 Conclusion

The primary purpose of this study is to understand customer satisfaction with Personal loans with reference to Yes Bank. This study found that employees in small ventures or private jobs mostly take personal loans. In most cases, customers are aware of the loan schemes through the telephone from the bank. Most of the respondents have taken loans of 5-10 lakhs. Furthermore, the reason for choosing the respective bank is that the respondent is already a bank customer due to competitive interest rates. Nowadays, Yes Bank takes only one week to sanction the loan. Most customers pay the EMI through Electronic Clearance Services (ECS). Most importantly, Yes Bank charges low for the foreclosure loan. The staff is also efficient in the bank, following RBI guidelines and providing a top-up loan facility with a low-interest rate.

5.1 Limitations

In this study, only exploratory factor analysis was done to summarize the variable. Further, confirmatory factor analysis can be done to delve into the study. The study was only conducted in the Delhi region, so the limited study area may affect the conclusions. Hence, this study can be extended by increasing the sample size, and subsequently, the analysis will be much more appropriate, and results may improve.

5.2 Future Scope

There is an enormous amount of scope for further research and application in the area of customer satisfaction with personal loans in the banking sector. Longitudinal analyses could be carried out to monitor changes in satisfaction over time and identify developing trends and possible areas for development. Furthermore, dividing up the clientele according to different criteria, such as behaviour or demographics, may offer a more in-depth understanding of the disparate needs of separate groups, guiding customized

banking tactics. Cross-cultural comparisons could clarify how cultural differences affect satisfaction levels and help banks provide effective services to diversified customers. In light of the current digital landscape, banks looking to maximize their digital services may find it crucial to examine how technology adoption affects consumer satisfaction with digital lending procedures. Additionally, examining how customer experience, contentment, and service quality interact may help identify areas for development that have the biggest effects on consumer perceptions.

References

[1] B. R. Jadhav, T. Sadija, 'A study on factors affecting the growth of personal loan as a product at HDFC Bank', ZENITH International Journal of Multidisciplinary Research, 8(11), pp. 99-111, 2018.

[2] J. Singh, 'Understanding the structure of consumers' satisfaction evaluations of service delivery', Journal of the Academy of Marketing Science, 19, pp. 223–244, 1991.

[3] L. Cai, K. F. Yuen, M. Fang, X. Wang, 'A literature review on the impact of the COVID-19 pandemic on consumer behaviour: implications for consumer-centric logistics', Asia Pacific Journal of Marketing and Logistics, 35(11), pp. 2682–2703, 2023.

[4] A. Anand, D. Aggrawal, M. Agarwal, 'Market assessment with OR applications', CRC press, 2019.

[5] C. McDaniel Jr, R. Gates, 'Marketing research', John Wiley & Sons, 2018.

[6] A. Anand, G. Bansal, 'Predicting customer's satisfaction (dissatisfaction) using logistic regression', International Journal of Mathematical, Engineering and Management Sciences, 1(2), pp. 77–88, 2016.

[7] A. Anand, S. Singhal, O. Singh, 'Optimal advertising duration for profit maximization', Journal of Management Analytics, 7(3), pp. 458–480, 2020.

[8] R. Grewal, G. L. Lilien, S. Bharadwaj, P. Jindal, U. Kayande, R. F. Lusch et al., 'Business-to-business buying: Challenges and opportunities', Customer Needs and Solutions, 2, pp. 193–208, 2015.

[9] J. R. Nielsen (Ed.), 'The correspondence principle (1918-1923)', Elsevier, 2013.

[10] P. Kotler, 'Marketing management', Pearson Education India, 2009.

[11] G. Bansal, A. Anand, V. S. S. Yadavalli, 'Predicting effective customer lifetime: an application of survival analysis for telecommunication

industry', Communications in Statistics-Theory and Methods, 49(10), pp. 2305–2320, 2020.

[12] T. H. Ho, Y. S. Zheng, 'Setting customer expectation in service delivery: An integrated marketing-operations perspective', Management Science, 50(4), pp. 479–488, 2004.

[13] S. Zaid, P. Patwayati, 'Impact of customer experience and customer engagement on satisfaction and loyalty: A case study in Indonesia', The Journal of Asian Finance, Economics and Business, 8(4), pp. 983–992, 2021.

[14] L. Aksoy, T. L. Keiningham, A. Buoye, B. Lariviere, L. Williams, I. Wilson, 'Does loyalty span domains? Examining the relationship between consumer loyalty, other loyalties and happiness', Journal of Business Research, 68(12), pp. 2464–2476, 2015.

[15] O. O. Thomas, 'Change management and its effects on organizational performance of Nigerian telecoms industries: Empirical Insight from Airtel Nigeria', International Journal of Humanities Social Sciences and Education, 1(11), pp. 170–179, 2014.

[16] D. Heckl, J. Moormann, 'How to design customer-centric business processes in the banking industry', Journal of Financial Transformation, 21, pp. 67–76, 2007.

[17] S. Wadhwa, M. Kumar, 'Effect of training on employee performance: investigating Indian banking sector', Journal of Graphic Era University, pp. 83–89, 2019.

[18] S. Aggarwal, 'Post-Merger Customer Satisfaction, Loyalty and Service Quality in the Banking Sector: A Bibliometric Analysis', IUP Journal of Bank Management, 21(4), pp. 80–97, 2022.

[19] S. Kapoor, R. Naik, 'Using SIR Model and Recurrence Formula to Predict the Spread of COVID-19 in Sambalpur: A Mathematical Study', Journal of Graphic Era University, pp. 231–256, 2021.

[20] A. Parasuraman, V. A. Zeithaml, L. L. Berry, 'Servqual: A multiple-item scale for measuring consumer perc', Journal of Retailing, 64(1), pp. 12, 1988.

[21] M. Hanif, S. Hafeez, A. Riaz, 'Factors affecting customer satisfaction', International Research Journal of Finance and Economics, 60(1), pp. 44–52, 2010.

[22] Singh, J., and Kaur, G. 'Customer satisfaction and universal banks: an empirical study', International Journal of Commerce and Management, 21(4), pp. 327–348, 2011.

[23] R. Chaudhary, Y. Junjhua, 'Customer perception and satisfaction towards Home Loans', International Journal of Research in Commerce and Management, 2(9), pp. 125–128, 2011.

[24] P. Dawar, 'A Study of factors affecting customer satisfaction in Present highly competitive banking industry', Asia Pacific Journal of Marketing & Management Review, ISSN, 2013.

[25] A. Jain, A. Kumar, M. K. Dash, 'Demographics and customer satisfaction: loan procedures variables', SCMS Journal of Indian Management, 11(2), pp. 35, 2014.

[26] Tao, F. (2014). Customer relationship management based on increasing customer satisfaction. International Journal of Business and Social Science, 5(5), pp. 256–263, 2014.

[27] P. K. Kapur, A. Anand, N. Sachdeva, 'Profit estimation for a product under warranty: An evaluation based on customer's expectation and satisfaction', International Journal of Reliability, Quality and Safety Engineering, 21(06), pp. 1450029, 2014.

[28] H. Begum, S. A. Khan, M. Aktar, 'Impact of factors affecting customer satisfaction towards corporate loan in the banking sector of Bangladesh', Journal of Science and Technology, 60(67), pp. 1994–0386, 2022.

[29] A. Anand, Chanchal, P. K. Kapur, Y. Tamura, 'Quantification of number of adopters: a study to showcase products-sold and products-in-use', International Journal of System Assurance Engineering and Management, pp. 1–13, 2023

[30] Tien, N. H., Anh, N., Dung, H., On, P., Anh, V., Dat, N., and Tam, B. Factors impacting customer satisfaction at Vietcombank in Vietnam. Hmlyan J. Econ. Bus. Manag, 2, pp. 44-51, 2021.

[31] Agag, G., Shehawy, Y. M., Almoraish, A., Eid, R., Lababdi, H. C., Labben, T. G., and Abdo, S. S. Understanding the relationship between marketing analytics, customer agility, and customer satisfaction: A longitudinal perspective. Journal of Retailing and Consumer Services, 77, pp. 103663, 2024.

Biographies

Divya received a bachelor's degree in Mathematics from the University of Delhi in 2020, and a master's degree in operational research from the University of Delhi in 2022, and she is currently pursuing a doctorate degree in Department of Operational Research from the University of Delhi.

Deepti Aggrawal is currently working as Assistant Professor at USME, Delhi Technological University, India. She obtained her PhD degree from Department of Operational Research, University of Delhi. She was Operations Manager in Axis Bank till she joined as a research scholar in the Department of Operational Research in 2011. Her Research areas include Marketing and Software Reliability. She is a life member of SREQOM and has publications in journals of national and international repute.

Adarsh Anand did his doctorate in the area of Innovation Diffusion Modeling in Marketing and Software Reliability Assessment. Presently he is working as an Associate Professor in the Department of Operational Research, University of Delhi (INDIA). He has been conferred with Young Promising Researcher in the field of Technology Management and Software Reliability by Society for Reliability Engineering, Quality and Operations Management (SREQOM) in 2012. He is a lifetime member of the Society for Reliability Engineering, Quality and Operations Management (SREQOM). He is also on the editorial board of International Journal of System Assurance and Engineering management (Springer). He has Guest edited several Special Issues for Journals of international repute. He has edited two books namely: "System Reliability Management (Solutions and Technologies)" and "Recent Advancements in Software Reliability Assurance" under the banner of Taylor and Francis (CRC – Press). He has publications in journals of national and international repute. His research interest includes modeling innovation adoption and successive generations in marketing, software reliability growth modelling and social media analysis.

The Scientific Temper (2024) Vol. 15 (2): 2349-2355
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2024.15.2.50
https://scientifictemper.com/

RESEARCH ARTICLE

Analysis of renewable energy and economic growth of Germany

Akanksha Singh, Nand Kumar*

Abstract

The study aims to investigate the effect of energy use, renewable energy consumption, nuclear energy consumption and air transport freight on CO_2 emissions in Germany. Auto-regressive distributed lag (ARDL) technique was used to establish the short and long-term correlations between the dependent (CO_2) and independent variables (Energy use, nuclear consumption, renewable consumption, air transport freight) on a time series data from 1970 to 2021. The results demonstrated the dependence of energy use on non-renewable energy sources by revealing a significant and positive relationship between short and long-term energy use and CO_2 emissions. Similar to energy use, nuclear energy use and air transport freight was also found to be positively correlated with CO_2 emissions. On the other hand, despite the fact that there is a strong negative connection between carbon dioxide emissions and renewable energy, the results reveal that the correlation between the two is not statistically significant. The study aims to investigate the effect of important economic factors on Germany's CO_2 emissions in order to move further in a sustainable manner. The data reveals an astonishing increase in the usage of nuclear energy for power generation between 1965 and 1985, which was followed by a fall in 2012 as a result of the Fukushima tragedy and growing social unease in Germany, which ultimately led to the phase-out of nuclear power. Nonetheless, BAU projections revealed that even with 200 nuclear power plants, CO_2 emissions would still be reduced to 278 by 2070, in part because of the additional power produced by other energy sources.

Keywords: CO_2 emissions, ARDL, Renewable energy consumption, Nuclear consumption, Air transport.

Introduction

One of the most important concerns of mankind today is perhaps climate change. This is clear from the fact that the world is no longer wasting time on theoretical debates and has turned its attention to concrete regulations and environmental laws. Therefore, it has become a matter of tremendous interest to spot and solve the challenges pertaining to reduction in carbon dioxide emission levels. Environmental degradation brought on by human activities that release CO_2 includes ecological problems, global warming, climate change and environmental pollution. According to BP statistics, the G7 nations contributed approximately 25% of global CO_2 emissions in 2019 and slightly less in 2020 to 23.2%. The world's most industrialized economies play a crucial role and their policies and practices are frequently imitated by other nations (B. Li and N. Haneklaus. 2022). We will examine the impact of macroeconomic variables on Germany's CO_2 emissions because it is one of the greatest energy consumers in the world and ranks third in the G-7 countries for overall CO_2 emissions, after the United States and Japan (M. Blesl. *et al*. 2006).

Environmental issues are having a growing impact on German energy policy. The government has set the objective of lowering greenhouse gas (GHG) emissions by at least 55% by 2030 and by 85% by 2050 compared to 1990 emission levels as part of the German Climate Action Plan 2050 (P. Markewitz., *et al*. 2019). The government has promoted initiatives for energy efficiency and renewable energy. The eco-tax was created in 1999 with the intention of promoting energy efficiency, renewable energy use, and conservation. The German Renewable Energy Act (EEG), which went into effect in 2000, established a feed-in tariff (FIT) system for all renewable technologies, which led to an astonishing increase in renewable energy capacity (C. Eui,

Department of Humanities, Delhi Technological University, New Delhi, Delhi, India.

***Corresponding Author:** Nand Kumar, Department of Humanities, Delhi Technological University, New Delhi, Delhi, India., E-Mail: nandkumar@dce.ac.in

How to cite this article: Singh, A., Kumar, N. (2024). Analysis of renewable energy and economic growth of Germany. The Scientific Temper, **15**(2):2349-2355.

Doi: 10.58414/SCIENTIFICTEMPER.2024.15.2.50

Source of support: Nil

Conflict of interest: None.

© The Scientific Temper. 2024

Received: 25/04/2024 **Accepted:** 10/05/2024 **Published**: 15/06/2024

C. Marcantonini, and A. D. Ellerman, 2013). However, only an energy shift is not sufficient to reduce GHG emissions significantly. In order to meet the ambitious ideals set by the German government, all sectors, including energy, industry, transport, agriculture and homes, will need to drastically reduce their GHG emissions. For a more comprehensive analysis, we have taken into account a number of sectors, including consumption of renewable energy, nuclear energy and air transportation along with energy use.

As per popular belief, clean energy fuels have a direct impact on the reduction of carbon dioxide levels through production as well as consumption. In this paper, we have emphasized on consumption, so renewable energy consumption is a significant regressor in our analysis. Renewable energy must be able to supply the necessary energy for economic activity and lessen environmental issues brought on by the use of fossil fuels (Y. Fang, 2011). It is worthwhile to mention that clean fuels not only require energy innovation at its pinnacle but also political and social commitment. Therefore, renewable energy consumption's impact may not be as direct and immediate as it seems. Trade is a crucial component of economic growth, which in turn implies that air travel is a crucial component as well. Despite its significance for the advancement of the nation and the creation of jobs, its unfavorable impact on GHG emissions is a serious concern. In fact, 915 million tonnes of CO_2 were created by flights globally in 2019, accounting for 12% of all transport-related CO_2 emissions. This leads us to the inclusion of air transport freight as an essential variable in the study. In an effort to build alternative sources of energy and to avoid the price fluctuations associated with oil import, nuclear power plants have been set up in several countries. Nuclear energy is proposed to offer a significant advantage in terms of reducing CO_2 emissions and curbing the effects of global warming (Y. Wolde-Rufael and K. Menyah, 2010). Hence, the nuclear consumption of Germany has also been taken into account in this study. The use of recent statistics, such as the BP statistics on G7 nation's CO_2 emissions, adds credibility and relevance to the discussion. The mention of Germany's position in global CO_2 emission provides a focused context for the study.

Review of Literature

Chang explored the relationships between energy use, carbon dioxide emissions and economic growth in China. The results of the study indicated that rising economic activity increases energy use and CO_2 emissions (C.C. Chang., 2010). A similar study on Russia's CO_2 emissions, energy use, and economic growth was analyzed by (H.T. Pao., *et al*. 2011). They used the causality test and the co-integration technique to analyze the data between the period of 1990 and 2007. Empirical findings indicated that emissions appear to be elastic in terms of energy use but inelastic in terms of output in the long run (A. Singh., *et al*. 2022) used ARDL analysis on data from 2007 to 2018. The study looks at how economic and environmental factors affect logistic performance in India. The logistics performance index (LPI) is found to be positively impacted by foreign direct investment (FDI) and negatively impacted by the use of fossil fuels, both in the short and long term. While greenhouse gas emissions show a positive correlation with LPI, GDP per capita has a negative correlation, raising worries about the environment. The global economy suffered as India's logistic performance rating fell from 35th to 44th in the world. The report makes suggestions for how to enhance logistics and explains why India

Some of the other studies focussed on factors different from energy use. Apart from energy consumption, Baek also investigated the effects of nuclear energy and income on CO_2 emissions in 12 major nuclear-generating countries. The research demonstrated that nuclear energy typically reduces CO_2 emissions (J. Baek, 2015). Related research by Al-mulai examined the impact of nuclear energy use on GDP growth and CO_2 emissions in 30 nations worldwide. But unlike the previous research, the study's results revealed that nuclear energy use has no long-term impact on CO_2 emissions, but it has a positive long-term impact on GDP growth (U. Al-mulali., 2014)

Some research also studied the relationships of renewable energy use (K. Menyah and Y. Wolde-Rufael, 2010) investigated the relationship between nuclear and renewable energy use and carbon dioxide emissions in the US from 1960 to 2007. The research showed no correlation connecting renewable energy to CO_2 emissions, but it discovered a unidirectional causality between the consumption of nuclear energy and CO_2 emissions. (S. Kumari and N. Kumar., 2023) concludes that while renewable energy boosts the green economy, helps agriculture, and lowers CO_2 emissions, it also highlights the need for better nuclear waste management to avoid environmental risks (A. Singh., *et al*. 2023). The study assesses GDP, nuclear energy consumption, and CO_2 emissions in the United States, France, and Japan from 1965 and 2020 in order to validate the environmental Kuznets curve (EKC) theory. It is demonstrated that nuclear energy greatly lowers CO_2, SO_2, and NOx emissions using the Fourier ARDL and FBTY causality tests. The report recommends boosting the generation of nuclear and tidal energy and making sure that radioactive waste is disposed of safely in order to reach the UN's 2050 net-zero carbon targets.

Similar research was conducted, which studied the causal association between renewable energy use and economic growth in Turkey from 1990 to 2010. The result revealed that there is a unidirectional causal relationship linking economic growth to renewable energy use (O. Ocal and A. Aslan, 2013). Sebri investigated related research for the BRICS countries

between 1971 and 2010 using a multivariate approach. The research showcased that based on the ARDL estimations, empirical data demonstrates that there are long-run equilibrium relationships between the variables (M. Sebri and O. Ben-Salha, 2014). Similar techniques like ARDL and VAR methodology was used to examine the relationship between CO_2 emissions, economic growth, energy consumption and renewable energy consumption in Germany for the years 1975 to 2014 and suggested that the government should create and implement appropriate support programmes to encourage investment in new renewable energy technology by (S. khoshnevis Yazdi and B. Shakouri, 2018).

Different factors from energy use, nuclear energy use, and renewable energy use were analysed in some of the other research. (R. Waheed., *et al*. 2017) Studied the impact of Pakistan's agricultural and forest production on carbon dioxide emissions in addition to renewable energy use. The research showcased that the use of renewable energy and forests have negative and significant effects on CO2 emissions, whereas agricultural production has a positive impact on CO_2 emissions. (Y.Cai., *et al*. 2018) Examined the nexus between clean energy usage, economic growth and CO_2 emissions. The result revealed that in Canada, France, Italy, the US, and the UK there was no correlation between real GDP per capita, clean energy consumption and CO_2 emissions. However, when real GDP per capita and CO_2 emissions are the dependent variables, co-integration occurs in Germany and when CO_2 emissions are the dependent variable, co-integration occurs in Japan.

Methodology

The study establishes a relationship between CO_2 emissions and energy use, renewable energy consumption, nuclear energy consumption and air transport freight over the period 1970–2021 in Germany.

For the purpose of this study, we have used secondary data that was acquired from the World Bank Development Indicators (WDI). Equation 1 establishes the relationship of CO_2 emission with energy use, nuclear consumption, renewable consumption and air transport freight in the context of Germany during the period 1970-2021.

$$\ln CO_2 = f(\ln \text{Energy}, \ln \text{Nuclear}, \ln \text{Renew}, \ln \text{Air}) \quad \ldots.(1)$$

The variables are described in Table 1 and in, the natural log, is used in the equation above.

$$\Delta \ln CO_2 = \alpha_0 + \sum_{i=1}^{n_1} \alpha_{1i} \Delta \ln ENERGY_{t-i} + \sum_{i=1}^{n_2} \alpha_{2i} \Delta \ln NUCLEAR_{t-i} + \sum_{i=1}^{n_3} \alpha_{3i} \Delta \ln RENEW_{t-i} + \sum_{i=1}^{n_4} \alpha_{4i} \Delta \ln AIR_{t-i} + \beta_1 \ln ENERGY_{t-1} + \beta_2 \ln NUCLEAR_{t-1} + \beta_3 \ln RENEW_{t-1} + \beta_4 \ln AIR_{t-1} + \mu_t \quad \ldots. (2)$$

In equation 2 the short-run relationship is represented by α_1 to α_4 whereas the long-run relationship is represented by β_1 to β_4 where α_0 is the drift component and μ_t is the error term and n_i is the optimal lag.

$$\Delta \ln CO_{2t} = \beta_0 + \sum_{i=1}^{n_1} \beta_{1i} \Delta \ln ENERGY_{t-i} + \sum_{i=1}^{n_2} \beta_{2i} \Delta \ln NUCLEAR_{t-i} + \sum_{i=1}^{n_3} \beta_{3i} \Delta \ln RENEW_{t-i} + \sum_{i=1}^{n_4} \beta_{4i} \Delta \ln AIR_{t-i} + \theta ECM_{t-1} + \mu_t \quad \ldots. (3)$$

In order to create an error correction model, the co-integration of the variables was first tested, whereas θ is the speed of adjustment of the long-run equilibrium.

The description of the variable is shown in Table 1 where ENERGY, NUCLEAR, RENEW and AIR are the independent variables affecting CO_2 emission. We discovered that the ARDL technique was best suited for our research since some variables were stationary at the first difference and some at a level.

Results and Discussion

A popular statistical test to determine whether a particular time series is stationary or not is the augmented Dickey-Fuller test (ADF test). In time series analysis, the unit root test is used to determine if a time series is stationary or not. The null hypothesis states that time series have a unit root, whereas the alternative hypothesis states that time series are stationary. Therefore, we require the probability values to be smaller than 0.05 in order to reject the null hypothesis. The outcome of the unit root test is shown in Table 2. The results proved the stationarity of the dependent variable CO_2 at a level as well as 1st difference, whereas the independent variables were stationary at 1st difference.

For the purpose of examining co-integration, Pesaran, Shin and Smith (PSS) created the ARDL bounds test. (A. Montenegro., 2019) The ARDL model is the most suitable for our study since it permits having a mixture of variables at a level and first difference. Another advantage that ARDL test offers is that it is relatively more efficient in the case of small and finite sample data sizes (M. Belloumi, 2014). The ARDL bound test is used to evaluate the long-term connection between the dependent and independent variables prior to performing ARDL co-integration. The upper bound and lower bound are the two essential values that the bound test

Table 1: List of variables

Variables	*Description*
CO_2	Carbon dioxide emissions (Million Tonnes)
Energy	Energy use (kg of oil equivalent per capita)
Nuclear	Nuclear consumption (EJ)
Renew	Renewable consumption (EJ)
Air	Air transport, freight (million ton-km)

Source: [WorldBank Database, 2020]

Table 2: ADF unit root test

Variables	*Level t-statistics*	*Level probability*	*1st difference t-statistics*	*1st difference probability*
CO_2	-4.278871	0.0070	-8.039970	0.0000
Energy	-3.318208	0.0748	-7.844101	0.0000
Nuclear	-2.700536	0.2408	-4.932591	0.0011
Renew	-1.581440	0.7860	-1.615187	0.0995
Air	-1.279795	0.8816	-9.082132	0.0000

Table 3: Auto-regressive distributed lag bond test for co-integration

Variable	*F statistics*	*Co-integration*	
F (CO_2, energy, nuclear, renew, air)	4.445385***		
Critical value (%)	1	5	10
Lower bound	3.29	2.56	2.20
Upper bound	4.37	3.49	3.09

*Significant at 10% level, **significant at 5%, ***significant at 1%

Table 4: Auto-regressive distributed lag short-run estimates

Variables	*Probability*	*t-statistic*	*Co-efficient*
Energy	0.0000***	7.733957	0.892228
Nuclear	0.1056*	1.655687	0.038417
Renew	0.1581	1.438348	0.039372
Air	0.0002***	4.112279	0.135124

*Significant at 10% level, **significant at 5%, ***significant at 1%

characterizes. The lower limit assumes that all variables are at level, while the upper bound presumes that all variables are at first different. The null hypothesis is rejected and co-integration is shown to exist if the upper bound value is less than the F statistic.

Table 3 displays the results of the ARDL bound test; the findings reveal that the dependent and independent variables exhibit considerable co-integration because the F statistic is significantly higher than both the upper bound and lower bound statistic.

The findings of the ARDL short-run test are shown in Table 4. Energy use (0<0.05) and air transport (freight) (0.002<0.05) are significant and have a positive correlation with CO_2 emission, whereas nuclear consumption also has a positive relation with CO_2 emission and is significant at 10% (0.10>0.05). On the other hand, renewable energy has no significant correlation with CO_2 emission in the short run.

Table 5 describes the results of the long-run ARDL test. Energy use, is statistically significant and is positively correlated with carbon dioxide emission. Similarly, nuclear energy is positively correlated with CO_2 emissions and is significant at 10%. Moreover, air transport (freight) is again significant and has a positive correlation with CO_2 emission. On the other hand, renewable energy, an important variable of our study, is not statistically significant and has a negative correlation with carbon dioxide emissions.

Table 6 illustrates the results of the diagnostics tests that were performed to validate our findings. The Jarque-Bera test, which has achieved widespread acceptability among econometricians, is one of the most well-known tests for determining the normality of regressors. (T. Thadewald and H. Büning., 2007) Since normality is the most common presumption when using statistical procedures, we applied the Jarque-Bera test, and the findings indicate that variables are normally distributed.

Table 5: Auto-regressive distributed lag long-run estimates

Variables	*Probability*	*t-statistic*	*Coefficient*
Energy	0.0046**	3.000507	0.639761
Nuclear	0.1056*	1.655687	0.038417
Renew	0.3594	0.927247	0.028450
Air	0.0002***	4.112279	0.135124

*Significant at 10% level, **significant at 5%, ***significant at 1%

Table 6: Diagnostic tests

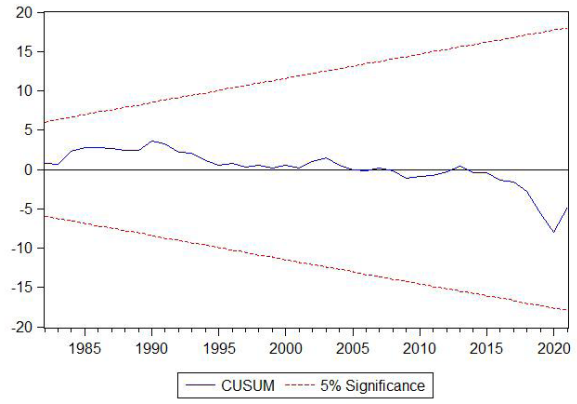
Jarque-Bera normality test JB stat	0.2346	0.8893
Ramsey RESET test, Log likelihood ratio	1.0537	0.2985
Breusch-Godfrey serial correlation, LM test	2.2290	0.3281
Heteroskedasticity Breusch-Pagan-Godfery test, Obs R-squared	15.3637	0.0814

Table 7: Granger causality test

Direction of causality	F-statistics	Probability
Energy - CO_2	1.76577	0.1827
CO_2 - Energy	6.13381	0.0044**
Nuclear - CO_2	0.61963	0.5427
CO_2 - Nuclear	4.48486	0.0167**
Renew - CO_2	1.76303	0.1832
CO_2 - Renew	2.27442	0.1146
Air - CO_2	2.59037	0.0861*
CO_2 - air	0.18345	0.8330
Nuclear - energy	4.08311	0.0235**
Energy - nuclear	2.11688	0.1322
Renew - energy	5.07704	0.0103**
Energy - renew	0.28825	0.7509
Air - energy	6.70181	0.0028***
Energy - air	0.96116	0.3902
Renew - nuclear	2.2363	0.1186
Nuclear - renew	2.90893	0.0649*
Air - nuclear	5.03021	0.0107**
Nuclear - air	1.21720	0.3056
Air - renew	5.13433	0.0098***
Renew - air	0.81349	0.4497

Homoscedasticity is a crucial presumption in statistical analysis. In order to determine if the data are homoscedastic or heteroscedastic, the heteroskedasticity Breusch-Pagan Godfrey test has been used; the results show that the dataset is homoscedastic.

Further Ramsey RESET test was used, and the outcome demonstrates that the model is free from specification error. Specification error means that at least one of the model's fundamental characteristics or presumption is erroneous.



*Significant at 10% level, **significant at 5%, ***significant at 1%

Figure 1: CUSUM test

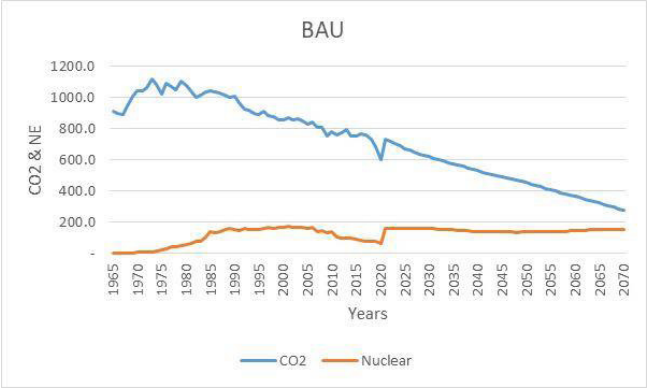


Figure 2: The business usual estimates

As a result, model estimation could produce inaccurate or deceptive results. Lastly, the Breusch-Godfrey Serial Correlation LM test demonstrated that our model's error component is not serially correlated.

Table 7 shows the results of the Granger causality test, the test is applied to time series data to determine whether there is any pronounced predictive power from one variable to the other. In essence, it aids in determining if a variable is helpful or not and whether it is capable of predicting other factors or not. It describes a one-to-one relationship between variables. As per our results, CO_2 - energy use, CO_2 - nuclear consumption, CO_2 - air, energy use - nuclear consumption, energy use - renewable energy consumption, energy use - air transport (Freight), renewable energy - nuclear consumption, nuclear consumption - air transport (Freight), renewable energy - energy use have unidirectional causal relationships.

Figure 1 demonstrates the result of the cumulative sum of the recursive residual stability test (CUSUM). Brown *et al*. invented the CUSUM test in 1975. It is a common statistical tool for analyzing and tracking structural changes in time series models (S. Otto and J. Breitung., 2022). The Figure 1 implies that the model is stable because our graph is within the bounds of 5% significance level.

BAU Estimates

The analysis a trend of sharp increase in the use of nuclear energy for power generation between the years 1965 to 1985. However, the trend started declining from the year 2012 mainly because of the 2012 Fukushima accident and the concerns in German society regarding nuclear accidents were deep, so the government had to completely phase out nuclear power generation. However, the BAU estimates from 2022 to 2070 show that even if the nuclear power generation would have maintained the 200 mark then also the CO_2 emission would have dropped to the mark of 278 by the year 2070. It's also because of the increase in power generation through other sources of renewable energy (Figure 2).

Conclusion

CO_2 emission is a major contributor to climate change and global warming. Substantial increase in CO_2 emission over the years has altered the proportions of our atmosphere and has given rise to global temperatures. The negative effects of this are currently being felt throughout the planet in the form of abrupt changes in weather patterns, extremely high temperatures, rising sea levels and melting ice caps. In this era of widespread recognition of the current issue, countries all over the world are issuing laws and taking precautions, but before one can be sure that these actions are effective, one must first determine whether or not they are even addressing the proper cause. In this paper, we analyzed whether or not the theoretically stated variables had an impact on Germany's CO_2 emissions. One can make some inferences from the empirical testing carried out in this paper.

The focus of the study is to find out whether CO_2 emissions are affected by energy use, renewable energy consumption, nuclear consumption, and air transport (freight) in Germany from the period 1970 to 2021. The results show a substantial and positive correlation between short and long-term energy use and CO_2 emissions, demonstrating the dependence of energy use on non-clean energy sources. The amount of energy used nationwide directly affects carbon dioxide emissions. Similarly there is a positive association between air transport (freight) and CO_2 emissions, indicating that the expansion of the aviation industry has contributed to increased emissions.

Consumption of nuclear energy is positively correlated with CO_2 emissions. It is obvious that Germany's CO_2 emissions will not decrease by switching to nuclear energy. As a result, the country's policy of using nuclear energy as a clean fuel has not been successful. On the contrary, we observe a rise in CO_2 emissions due to other factors such as energy use, industry, population, etc. Although there is a negative correlation between carbon dioxide emissions

and renewable energy, which is a significant variable in our analysis, the test findings show that this correlation is not statistically significant. This suggests that just implementing policies to increase the use of renewable energy has little effect on carbon dioxide emissions. The aforementioned two findings are quite important since they show that, in the case of Germany, the energy industry has no impact at all on CO_2 emissions. Significant CO_2 emissions are not being caused by energy generation or consumption. Other economic sectors like transportation, industrialization, etc, instead influence CO2 emissions. As a result, the nation needs to change its focus on the source of CO_2 emissions. In a complicated economic system, it is exceedingly difficult to comprehend the underlying causes of CO_2 emissions. We can never be certain that our efforts to conserve the environment will be successful if we don't understand the true cause. As a result, ongoing efforts are made in this area, and our work has been essential in understanding some of the real variables influencing Germany's CO_2 emissions. The data demonstrates a strong rise in the usage of nuclear energy for power generation between 1965 and 1985, which was followed by a fall in 2012 as a result of the Fukushima tragedy and growing social unease in Germany, which ultimately led to the phase-out of nuclear power. Due in part to enhanced power generation from other renewable energy sources, BAU projections show that even with 200 units of nuclear power generation, CO_2 emissions will still be reduced to 278 by 2070. Further research can be done on managing the waste disposable generated by nuclear energy.

References

Al-Mulali, U. (2014). Investigating the impact of nuclear energy consumption on GDP growth and CO2 emission: A panel data analysis. *Progress in Nuclear Energy*, *73*, 172-178.

Baek, J. (2015). A panel co-integration analysis of CO2 emissions, nuclear energy and income in major nuclear generating countries. *Applied Energy*, *145*, 133-138.

Blesl, M., Das, A., Fahl, U., & Remme, U. (2007). Role of energy efficiency standards in reducing CO2 emissions in Germany: An assessment with TIMES. *Energy policy*, *35*(2), 772-785.

Belloumi, M. (2014). The relationship between trade, FDI and economic growth in Tunisia: An application of the autoregressive distributed lag model. Economic systems, 38(2), 269-287.

Chang, C. C. (2010). A multivariate causality test of carbon dioxide emissions, energy consumption and economic growth in China. *Applied Energy*, *87*(11), 3533-3537.

Cai, Y., Sam, C. Y., & Chang, T. (2018). Nexus between clean energy consumption, economic growth and CO2 emissions. Journal of cleaner production, 182, 1001-1011.

Economics, I. A. T. A., & Airbus, B. (2012). ATAG Beginner's Guide to Aviation Efficiency. *IPCC, ICAO, United Kingdom Department for Transport, Aviation: benefits beyond borders, Airbus, Boeing, ATAG Beginner's Guide to Aviation Efficiency, the Intergovernmental Panel on Climate Change (IPCC), IATA, ATAG, BBC News, AERO modelling system, Qantas retrieved from http:// www. atag. org/facts-and-figures. html*.

Fang, Y. (2011). Economic welfare impacts from renewable energy consumption: The China experience. *Renewable and sustainable energy Reviews*, *15*(9), 5120-5128.

https://www.google.com/search?q=https%3A%2F%2Fdata. world bank.org%2F+&sca_esv=d669033b40b15054& source=hp&ei= yHZNZv2dCN2C2roPgryE6Ao&iflsig= AL9hbdgAAAAAZk2E2BDmwYFq2wbUzCyrAULU8iBnQY7Q & ved=0ahUKEwj9od2NuKCGAxVd gVYBHQIeAa0Q4 dUDCA0&uact = 5&oq=https%3A%2F%2Fdata.world bank.org %2F+&gs_lp= Egdnd3 Mtd2l6IhxodHR wczov L2 Rhd GEud29ybGRiYW5 rLm9yZy8gMgYQABgNGB4 yCBAAGAoYDRgeMggQABgKGA0 YHjIGEAAYDRge Mg YQABg WGB4 yBhAAGBYYHjI GEAAYDRgeMg YQA Bg WGB4yChAAG AgY ChgNG B4yCBAAGA0 YHhgPSI gFUABYAH AAeACQAQCYAag BoAGoAaoBAz AuM bgBA8 gBAPgBAvgBAZg CAaACzAG YAwCSBwMyL TGgB78 M&sclient = gws-wiz

khoshnevis Yazdi, S., & Shakouri, B. (2018). The renewable energy, CO2 emissions, and economic growth: VAR model. *Energy Sources, Part B: Economics, Planning, and Policy*, *13*(1), 53-59.

Li, B., & Haneklaus, N. (2022). Reducing CO2 emissions in G7 countries: The role of clean energy consumption, trade openness and urbanization. *Energy Reports*, *8*, 704-713.

Montenegro, A. (2019). The Ardl Bounds Co-integration Test: Tips for Application and Pretesting. *Available at SSRN 3425994*.

Markewitz, P., Zhao, L., Ryssel, M., Moumin, G., Wang, Y., Sattler, C., ... & Stolten, D. (2019). Carbon capture for CO2 emission reduction in the cement industry in Germany. *Energies*, *12*(12), 2432.

Menyah, K., & Wolde-Rufael, Y. (2010). CO2 emissions, nuclear energy, renewable energy and economic growth in the US. *Energy policy*, *38*(6), 2911-2915.

Marcantonini, C., & Ellerman, A. D. (2013, May). The cost of abating CO2 emissions by renewable energy incentives in Germany. In *2013 10th International Conference on the European Energy Market (EEM)* (pp. 1-8). IEEE.

Ocal, O., & Aslan, A. (2013). Renewable energy consumption–economic growth nexus in Turkey. *Renewable and sustainable energy reviews*, *28*, 494-499.

Otto, S., & Breitung, J. (2023). Backward CUSUM for testing and monitoring structural change with an application to COVID-19 pandemic data. *Econometric Theory*, *39*(4), 659-692.

Pao, H. T., Yu, H. C., & Yang, Y. H. (2011). Modeling the CO2 emissions, energy use, and economic growth in Russia. *Energy*, *36*(8), 5094-5100.

S. Kumari and N. Kumar., "Analysing the sustainability of renewable energy in France" 2023 Eco. Env. & Cons. 29 (November Suppl. Issue) : 2023

Singh, A., Lal, S., Kumar, N., Yadav, R., & Kumari, S. (2023). Role of nuclear energy in carbon mitigation to achieve United Nations net zero carbon emission: evidence from Fourier bootstrap Toda-Yamamoto. *Environmental Science and Pollution Research*, *30*(16), 46185-46203.

Singh, A., Kumari, S., Singh, B., & Kumar, N. (2022). Investigating the economic and environmental sustainability of logistic operations in India using ARDL procedure. *Advanced Production and Industrial Engineering*, 81-90.

Sebri, M., & Ben-Salha, O. (2014). On the causal dynamics between economic growth, renewable energy consumption, CO2

emissions and trade openness: Fresh evidence from BRICS countries. *Renewable and Sustainable Energy Reviews*, *39*, 14-23.

Thadewald, T., & Büning, H. (2007). Jarque–Bera test and its competitors for testing normality–a power comparison. *Journal of applied statistics*, *34*(1), 87-105.

Waheed, R., Chang, D., Sarwar, S., & Chen, W. (2018). Forest, agriculture, renewable energy, and CO2 emission. *Journal of Cleaner Production*, *172*, 4231-4238.

Wolde-Rufael, Y., & Menyah, K. (2010). Nuclear energy consumption and economic growth in nine developed countries. *Energy economics*, *32*(3), 550-556.

Physica Scripta

INCLUSIVE PUBLISHING
TRUSTED SCIENCE

PAPER

Analytical modeling of nucleation and growth of graphene layers on CNT array and its application in field emission of electrons

To cite this article: Shruti Sharma and Suresh C Sharma 2024 *Phys. Scr.* **99** 075918

View the article online for updates and enhancements.

You may also like

- Self-powered boost-converter for power optimisation and piezo garden lights
Mahesh Edla, Deguchi Mikio, Iman Izadgoshasb et al.
- Point and counterpoint between Mathematical Physics and Physical Mathematics
P G L Leach and M C Nucci
- Viable Approach for Forming Uniform Polymer Nanocomposites with Ultrahigh Filler Loading
Chia-Yun Hsieh and Kenneth K.S. Lau

This content was downloaded from IP address 14.139.251.98 on 11/06/2024 at 10:55

Physica Scripta

PAPER

Analytical modeling of nucleation and growth of graphene layers on CNT array and its application in field emission of electrons

Shruti Sharma and Suresh C Sharma

Department of Applied Physics, Delhi Technological University, Shahbad Daulatpur, Bawana Road, Delhi, 110042, India

E-mail: suresh321sharma@gmail.com

RECEIVED
4 January 2024

REVISED
5 May 2024

ACCEPTED FOR PUBLICATION
22 May 2024

PUBLISHED
6 June 2024

Keywords: CNT –PGL hybrid, PECVD, C2H2 flow, catalyst film thickness, substrate temperature, cooling rate, precipitation

Supplementary material for this article is available online

Abstract

Carbon Nanotube (CNT) arrays and graphene have undergone several investigations to achieve efficient field emission (FE) owing to CNT's remarkable large aspect ratio and graphene's exceptional FE stability. However, when dense CNT arrays and planar graphene layers were used as field emitters, their field enhancement factor reduced dramatically. Therefore, in this paper, we numerically analyze the growth of a dense CNT array with planar graphene layers (PGLs) on top, resulting in a CNT-PGL hybrid and the associated field enhancement factor. The growth of the CNT array is investigated using Plasma Enhanced Chemical Vapor Deposition (PECVD) chamber in C_2H_2/NH_3 environment with variable C_2H_2 flow, Ni catalyst film thickness, and substrate temperature followed by PGL precipitation on its top at an optimized cooling rate and Ni film thickness. The analytical model developed accounts for the number density of ions and neutrals, various surface elementary processes on catalyst film, CNT array growth, and PGLs precipitation. According to our investigation, the average growth rate of CNTs increases and then decreases with increasing C_2H_2 flow rate and catalyst film thickness. CNTs grow at a faster rate when the substrate temperature increases. Furthermore, as the chamber temperature is lowered from 750 °C to 250 °C in N_2 environment and Ni film thickness grows, the number of the graphene layers increases. The field enhancement factors for the CNT array and hybrid are then calculated based on the optimal parameter values. The average height of the nanotubes, their spacing from one another, and the penetration of the electric field due to graphene coverage are considered while computing the field enhancement factor. It has been found that adding planar graphene layers to densely packed CNTs can raise its field enhancement factor. The results obtained match the current experimental observations quite well.

1. Introduction

Carbon nanomaterials having hexagonally arranged sp^2 hybridized carbon atoms are actively used in various application domains. They are divided into three categories based on their dimensional shape: one-dimensional (1D) i.e. carbon nanotube (CNT), two-dimensional (2D) i.e. graphene and three-dimensional (3D) i.e. CNT-graphene hybrid. Carbon nanomaterials have the following set of properties: high electrical and thermal conductivity [1–3], improved surface area, high aspect ratio, field emission stability, flexibility, and low voltage requirement for electron extraction from its surface [4]. They are employed as electron field emitters [5, 6] in x-ray generators [7], flat panel displays [8], microwave amplifiers [9], in electrodes for lithium-ion batteries for enhancing electron and Li-ion transport [10, 11], in supercapacitors [12, 13], in biosensors where CNT's are functionalized with biomolecules [14]. Moreover, they are used as field emitters mounted on hall effect thrusters and in space electric propulsion systems to increase channel wear resistance [15–17].

CNT- Graphene hybrid shows exemplary field emission (FE) owing to the adorable properties of both CNT and graphene, i.e., high aspect ratio, field emission stability and sharp edges of graphene [4, 18, 19].

© 2024 IOP Publishing Ltd

Table 1. Neutrals and ions taken into consideration [34–37].

Neutrals			Ions		
Type j	Type k	Type l	Type j	Type k	Type l
C_2H_2, CH_4, CH_3, C_2H_3, C_3H_4, C_4H_2	H, H_2, H_3	NH_3, N_2, NH_4, HCN,	$C_2H_2^+$, CH_4^+, CH_3^+, $C_2H_3^+$, $C_4H_2^+$	H^+, H_2^+, H_3^+	NH_3^+, NH_4^+, HCN^+, CN^+

PECVD (Plasma Enhanced Chemical Vapor Deposition) is the most widespread and versatile method for fabricating carbon nanostructures. In PECVD, the electrostatic force from the applied electric field causes ionization, generating electrons and ions that participate in various reactions (excitation, dissociation), forming radicals. Because of this, the technique can be operated at low temperatures without damaging the substrate and forming vertically aligned carbon nanostructures [20]. The nucleation of the carbon nanostructure in PECVD is controlled by the plasma species (ions, electrons, and neutral atoms) generated, catalyst film thickness, precursor gas flow rate, and plasma control parameters (gas pressure, applied plasma power, bias) [21–24].

Choi *et al* [25] reported that catalyst nanoparticles are formed by plasma pretreatment of the catalyst surface prior to hybrid growth in an optimum growth environment within a PECVD chamber. Extensive studies have shown that the graphene layer number, diameter, and crystallinity of CNT are determined by catalyst film thickness and nanoparticle size [26–28]. It is also scrutinized that CNT array morphology can be directed by properly fixing the growth time, substrate temperature, and gas ratio [29, 30].

Chhowalla *et al* [30] and Bell *et al* [31] studied the effects of the C_2H_2/NH_3 feedstock flow ratio on CNT array growth. It was validated that when the concentration of C_2H_2 is high (>50%), pyramidal structures rather than vertical nanotubes emerge. Choi *et al* [25] and Yu *et al* [26] demonstrated the effects of cooling rate on the yield of graphene from Ni nanoparticles and flat Ni substrate, and proposed an optimal cooling rate for C precipitation. Shurman *et al* [27] simulated the formation of CVD graphene on a thin Ni film at various growth parameters such as growth time, temperature, and Ni film thickness.

Felicitous designing of the CNT array and controlling its dimensions make it worthy as a field emitter in FE devices. Bonard *et al* [32] investigated the effects of geometric configurations (density, height, radius, and distance between CNTs) on the FE properties of CNTs, and found that high-density CNT films are not a prerequisite for optimal field emission. Thapa *et al* [29] investigated that single CNTs or less crowded CNT bundles provide better field emission than crowded CNT bundles. In comparison to pure CNT and graphene films, Kaur *et al* [33] and Hong *et al* [19] observed increased field emission from reduced graphene oxide (rGo) deposited CNT films over Ni foams as well as a decrease in turn on electric field and threshold electric field for graphene embedded CNTs.

Motivated by the aforementioned studies, the author has tried to model and numerically analyze the growth of the CNT-Planar graphene layer (PGL) hybrid. The author determined the optimal dimensions of the hybrid for enhanced field emission by analyzing the favorable C_2H_2 flow rate, Ni film thickness, substrate temperature, and cooling rate.

2. Model

This section works out the kinetics of plasma species and the growth of the CNT-PGL hybrid in C_2H_2/NH_3 plasma using fundamental assumptions and equations. Using the DC-PECVD process, a hybrid of PGLs and CNT arrays is designed over a Silicon (Si) substrate, covered with Ni film as a catalyst. Table 1 lists basic reactants such as ions and neutrals of C_2H_2, NH_3, and H_2 that comprise C_2H_2/NH_3 plasma [34].

In the PECVD technique, when a negatively biased substrate surface with a catalyst film is exposed to a reactive plasma, an electric field arises between the bulk plasma and the substrate surface because of charge separation. The charge separation region is known as the plasma sheath. It is composed of both neutral molecules and positive ions. The electrostatic force arising from the plasma sheath accelerates ions toward the catalyst film surface, i.e., in the vertically downward direction (z-axis) [38], while neutrals undergo Brownian motion across the sheath. The plasma sheath equations (equations (1)–(5)) describe the plasma sheath. The parameters needed for hybrid growth namely the number density and speed of ions, as well as the number density of electrons within the plasma sheath are computed using the following continuity (equations (1) and (2)), momentum balance equations (equations (3) and (4)) for ions and electrons in plasma sheath respectively.

$$\frac{1}{n_{it}}\frac{\partial(n_{it}\nu_{it})}{\partial z} = \tau, \tag{1}$$

Table 2. Symbols with their meanings used in equations (1)–(5).

Symbol	Meaning
m_{it},	Mass of positive ion (t corresponds to type j, k, l)
n_{it},	Number density of positive ions (t corresponds to type j, k, l)
T_{it},	Temperature of positive ions (t corresponds to type j, k, l)
v_{it}	Velocity of positive ions (t corresponds to type j, k, l)
m_e,	Mass of electron
N	Number density of electrons
T_e	Temperature of electrons
v_e	Velocity of electrons
τ	Ionization frequency of neutrals by electrons
$l_\eta = (\sigma_\eta N_u S_\eta)$	Collision frequency computing ions and electrons collisions with the neutrals
η	Refer to electrons and t (type j, k, l) ions
σ_η	Collision cross-section
S_η	Average speed of electrons and positive ions
N_u	Total neutral species number density.
$\beta_{it} = n_{it}/n$	tth ions to electrons number density ratio
ϕ	Electric sheath potential

$$\frac{1}{n}\frac{\partial(nv_e)}{\partial z} = \tau, \tag{2}$$

$$m_{it}v_{it}\frac{\partial v_{ik}}{\partial z} = -e\frac{\partial \phi}{\partial z} - m_{it}\iota_{it}v_{it} - \frac{T_{it}}{n_{it}}\frac{\partial n_{it}}{\partial z}, \tag{3}$$

$$m_e\iota_e v_e = e\frac{\partial \phi}{\partial z} - \frac{T_e}{n}\frac{\partial n}{\partial z}, \tag{4}$$

$$\frac{d^2\phi}{dz^2} = 4\pi e\left(\sum_t \beta_{it}n_{it} - n\right), \tag{5}$$

Poisson's equation (equation (5)), determines the potential distribution within the plasma sheath [39, 40]. Table 2 provides the interpretation of each symbol used in equations (1)–(5).

Ultimately, the sheath equations are resolved by applying the following boundary conditions at plasma sheath interface i.e. at z = 0:

$$\phi = 0,\ \frac{d\phi}{dz} = -\frac{T_e}{e\lambda},\ v_e = 0,\ v_{it} = v_{ito},$$

where $v_{ito} = \sqrt{\frac{T_e}{m_{it}}}$ is the ion –acoustic speed, $\lambda = \frac{T_i}{\sigma_\eta p_o}$ is the mean free path of ion, p_0 is total gas pressure, respectively [39].

Figure 1 illustrates the CNT-PGL hybrid's inculpated growth processes in a plasma environment. Initially under applied plasma power, during pre-etching of the Ni- catalyst film, the dissociation of NH_3 produces energetic electrons, hydrogen and ammonia ions, and neutrals. These plasma species strike at the Ni film due to an induced electric field (E) between the negatively biased substrate and bulk plasma to form Ni catalyst nanoparticles. These nanoparticles serve as nucleation sites for CNT array growth in the C_2H_2/NH_3 plasma environment [41]. The incoming fluxes of positive and neutral ions from the plasma are exposed to the catalyst nanoparticles, which aid in the formation of CNTs. Carbon and hydrogen species in the plasma dissociate and adsorb, across the catalyst surface through surface reactions. Adsorbed carbon atoms diffuse over the catalyst nanoparticle's surface and in bulk up to a diffusion length $\Lambda \approx \sqrt{(D_a t')}$ where t' is the diffusion time. Eventually, carbon atoms in the catalyst bulk begin to precipitate beneath the catalyst surface to direct the tip growth –mode (i.e. lift up of catalyst from the substrate surface) of the CNTs in the CNT array [42], in conjunction with etching by adsorbed hydrogen atoms and incoming nitrogen flux. The electrostatic force present in the plasma sheath causes CNTs to grow vertically [43]. While the array grows, the overall gas pressure is kept at 7 Torr.

After the CNT array growth, C_2H_2/NH_3 flow is terminated, plasma is turned off and the chamber is cooled to 250 °C under N_2 flow with total gas pressure 22 Torr. The cooling rate is assumed linear [25]. Three cooling rates are taken into consideration: 5 °C s^{-1}, 15 °C s^{-1}, and 25 °C s^{-1}. During the cooling period, carbon atoms begin to precipitate out of Ni nanoparticles with a driving force (F) due to the supersaturation of diluted carbon. These carbon atoms form tiny nuclei that expand and meet with one another leading to graphene growth. These

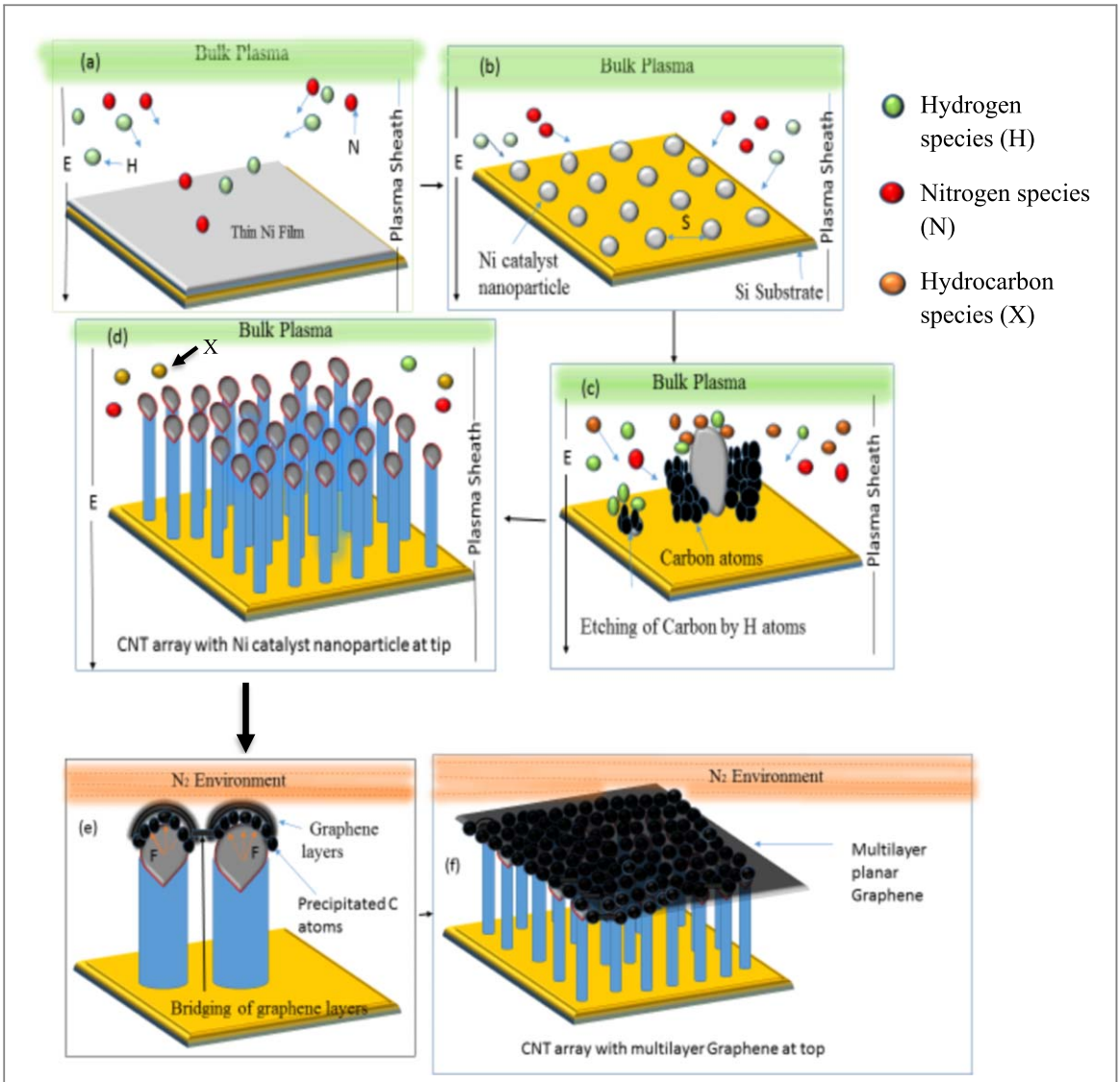


Figure 1. Schematic showing the (a) Pre-etching of the thin Ni film by incident plasma species in NH_3 plasma. (b) Cloven of Ni film into nanoparticles in NH_3 plasma. (c) Etching of carbon on Ni catalyst nanoparticle along with dissociation and adsorption of hydrocarbon and hydrogen species on catalyst nanoparticle surface in C_2H_2/NH_3 plasma. (d) Growth of CNT array via tip growth mode. (e) Precipitation of carbon atoms with driving force (F) to form graphene layers on Ni nanoparticle surface and bridging among graphene layers in N_2 environment. (f) CNT array with multilayer planar graphene at top.

graphene nuclei unite to form a planar graphene sheet by bridging the gaps across the Ni nanoparticles located at the tip of CNTs [25, 44–46]. As a result, a hybrid of CNT and PGL is formed.

2.1. Presumptions considered for simplified modelling [21, 47, 48]

(a) Maxwellian distribution of electrons and ions is considered.

(b) Ion—surface interaction is assumed to be coulomb interaction.

(c) For the sheath to perform its function and repel electrons, the potential must be monotonically decreasing with increase in z. This will occur if $n_{it}(z) > n(z)$ for all z in the sheath.

(d) Catalyst nanoparticle and substrate temperature is considered to be the same

2.2. Precipitation of carbon

Carbon atoms diffuse into Ni forming Ni-C bonds owing to its high solubility in Ni. When the chamber is cooled to a reasonably lower temperature, graphene formation occurs from diluted carbon in the Ni nanoparticles [49, 50]. When the carbon concentration in Ni reaches saturation during cooling, a Ni-C supersaturated solid solution is formed. Consequently, due to the difference in saturation while cooling, a driving force (F) comes into play, diffusing out carbon atoms from Ni bulk to surface and leading to carbon precipitation, following

Arrhenius law. According to Arrhenius law, the dependence of carbon out-diffusion from Ni bulk on temperature is given by

$$D_a = D_o \exp\left(\frac{-U_{voc}}{k_B T_s}\right), \tag{6}$$

where, D_a is bulk diffusion coefficient, $D_o = \frac{D_{Ni} f}{2\pi}$, U_{voc} (=1.74 eV) activation energy for diffusion of carbon atoms in Ni bulk [27, 45], T_s is the catalyst surface temperature , f is thermal vibration frequency ($\approx 10^{13}$Hz).

Let carbon concentration in Ni bulk along z direction, perpendicular to the surface be C_{bulk}(z, t) where t is time. According to Fick's Law of diffusion, the rate of change of concentration at a point within Ni bulk due to diffusion is given by

$$\frac{\partial C_{bulk}}{\partial t} = D_a \nabla^2 C_{bulk}, \tag{7}$$

where, t represents time and C_{bulk} (z, t) is estimated by

$$\frac{C_{bulk}(z,\ t) - C_o}{C_1 - C_o} = erf\left(\frac{z}{2\sqrt{D_a t}}\right), \tag{8}$$

The following boundary conditions are used to compute the C_{bulk} (z, t).

C_{bulk} (z, t) (z > 0, t = 0) = C_1 (Initial bulk concentration in Ni catalyst nanoparticle), C_{bulk} (z, t) (z = 0, t) = C_o (bulk concentration adjacent to the Ni catalyst nanoparticle surface).

The number of planar graphene layers (N_{pg}) obtained at time t is then approximated to

$$N_{pg} \approx \left(\frac{\frac{D_a(C_1 - C_0)}{\sqrt{\pi D_a t}}\frac{V_{Ni}\rho_{Ni}}{M_{Ni}}}{N_a}\right), \tag{9}$$

where, N_a denotes surface concentration of atoms in one graphene layer $\left(\approx 3.8 \times 10^{19}\frac{atoms}{m^2}\right)$. [27, 44, 51, 52], V_{Ni}, ρ_{Ni} (=8.96 g cm^{-3}), and M_{Ni} (=58.69 amu) are the volume, density and mass of Ni catalyst nanoparticle.

2.3. Catalyst nanoparticles formation

During the pre-etching, owing to energy transfer (i.e. input power (P)) by plasma species, the catalyst film gets heated up and segregated. This is the result of multiple energy fluxes contributing viz power transfer due to: striking of plasma species, recombination reactions, chemical etching, incident ion sputtering, and energy loss to surroundings. Therefore the time-varying energy balance equation on Ni film in NH_3 plasma is,

$$P = \frac{\rho_{Ni} C_s T_s}{6}\frac{\partial \pi D_{Ni}^3}{\partial t} = \left[\sum_t^{k,l}[\xi_{it}\varepsilon_{it} + \xi_t\varepsilon_t + \xi_e\varepsilon_e]\right] + \left[\frac{3}{2}k_B T_s\sum_t^{k,l}(1 - \gamma_{it})\xi_{it}\right] + [(\upsilon_o \sigma_{ads} F_k U)A] + \left[\left(\sum_t^{k,l} F_{it}\, Y_I(U_i)(1 - \theta)A\right)\right] - [\sigma A(\eta_s T_s^4 - \eta_{surr} T_{surr}^4)], \tag{10}$$

where, ξ_{it}, ξ_e, ξ_t denotes the collection current at the surface of Ni catalyst film due to positively charged ions of type t (k, l), electrons, and neutrals of type t (k, l) and are defined as,

$$\xi_{it} = \left[n_{it}(z)A\left(\frac{k_B T_{it}}{2\pi^2 m_{it}}\right)^{0.5}\left(\sqrt{\frac{4}{\pi}}\left(\frac{eV_c}{\pi k_B T_{it}}\right)^{0.5} + \exp\left(\frac{eV_c}{k_B T_{it}}\right) erfc\left(\frac{eV_c}{k_B T_{it}}\right)^{0.5}\right)\exp\left(\frac{-eU_c}{k_B T_s}\right)\right], \tag{10a}$$

$$\xi_e = n(z)A\left(\frac{k_B T_e}{2\pi^2 m_e}\right)^{0.5}\exp\left[\frac{eU_c}{k_B T_s} + \frac{eV_C}{k_B T_e}\right], \tag{10b}$$

$$\xi_t = n_t(z)A\left(\frac{k_B T_t}{2\pi^2 m_t}\right)^{0.5}, \tag{10c}$$

Description of other terms used in equations (10)–(10c) is given in table 3.

The left-hand side of equation (10) accounts for the power transfer to Ni film and formation of Ni nanoparticle of diameter (D_{Ni}) with time.

The first term on the right-hand side in equation (10) represents the energy accumulated by striking ions, neutrals, and electrons over the surface. The second term represents energy transfer via neutral formation on the

Table 3. Description of terms in equations (10)–(10*c*).

Terms	Description
$n_{it}(z) = n_{it0}\left(1 - \frac{2e\Phi(z)}{m_{it}v_{ito}^2}\right)$	Spatial dependence of ions within plasma sheath.
$n(z) = n\exp\left(\frac{\lvert e\rvert\ \Phi(z)}{k_B T_e}\right)$	Spatial dependence of electrons within plasma sheath.
$\Phi(z) = \Phi_0\exp\left(\frac{-\lvert z\rvert}{\lambda}\right)$	Positional dependent potential within plasma sheath
$F_t = \left(\frac{n_t v_{thst}}{4}\right)$	Flux of impinging neutrals on the catalyst surface [53].
$F_{it} = \left(n_{it}\left(\frac{k_B T_{it}}{m_{it}}\right)^{0.5}\right)$	Flux of impinging ions on the catalyst surface [53].
$\varepsilon_\varsigma = \left(\left(\frac{2 - qH_\varsigma}{1 - qH_\varsigma}\right) - qH_\varsigma\right)k_B T_\varsigma$	Average energy collected at the surface of Ni catalyst film via ς (ions, neutrals of type t and electrons), $H_\varsigma = \left(\frac{U_B}{k_B T_s}\right)$, U_B is the energy barrier at the catalyst film surface [54].

Table 4. Symbols with their meanings used in equations (10)–(10*c*).

Symbol	Meaning
C_s(=0.104 cal g^{-1} $°C^{-1}$)	Specific heat of Ni catalyst film
A	Area of Ni catalyst film
$\gamma_{it}(\approx 1)$	Sticking coefficient of positive ions of type t (j, k, l)
σ_{ads}(=6.8 × 10^{-16} cm^2)	Cross-section for reaction with atomic hydrogen
υ_o(≈10^{15} cm^{-2})	Number of adsorption sites per unit area
η_{surr}(=1) and η_s(=0.12)	Environment emissivity and Ni catalyst film emissivity
σ(5.67 × 10^{-5} *ergs* $cm^{-2}s^{-1}K^{-4}$)	Stefan's constant
$Y_I(U_i)$	Sputtering yield, dependent on ion energy [55]
V_C	Potential at surface of Ni catalyst film
Uc = φ_o	Substrate potential
k_B	Boltzmann constant
θ = 0.01	Total surface coverage
U (=3.74 eV for Ni)	Binding energy of material surface [53]
T_S and T_{surr}	Substrate temperature and temperature of plasma environment
ν_{thst}	Thermal velocity of neutral species.

surface. Energy transfer from chemical etching and incident ion sputtering is represented by the third and fourth terms. The last term accounts for energy loss from the surface to the surroundings.

Table 4 provides the interpretation of each symbol used in equations (10)–(10*c*).

2.4. Charging of CNT-PGL hybrid

Charge (Z') is evolved over the surface of the CNT- PGL hybrid because of the augmentation of positive charges and electrons, and is computed through equation,

$$\dot{Z'} = \sum_{w=1}^{r}\left(\sum_{t}^{j,k,l}\xi_{it}^{tipw} + \sum_{t}^{j,k,l}\xi_{it}^{curw} + \sum_{t}^{j,k,l}\xi_{it}^{pgw} - \Upsilon_e(\xi_e^{tipw} + \xi_e^{curw} + \xi_e^{pgw})\right), \tag{11}$$

where, r represents number of CNTs , ξ_{it}^{tipw}, ξ_{it}^{curw}, ξ_{it}^{pgw} represents the ion collection current at w^{th} - CNT tip, curved surface and PGL and are defined as,

$$\xi_{it}^{tipw} = \left[\frac{\pi D_w^2}{4}\left(\frac{8k_B T_{it}}{\pi m_{it}}\right)^{\sqrt{0.25}} n_{it}(z)(1 - Z_{CNT}\Psi_V)\exp\left(\frac{-eU_c}{k_B T_s}\right)\right], \tag{11a}$$

$$\xi_{it}^{curw} = \left[n_{it}(z) \frac{D_w}{2} L_w \left(\frac{2\pi k_B T_{it}}{m_{it}} \right)^{\sqrt{0.25}} \left(2 \left(\frac{eV_{cur}}{\pi k_B T_{it}} \right)^{\sqrt{0.25}} + \exp\left(\frac{eV_{cur}}{k_B T_{it}} \right) erfc \left(\frac{eV_{cur}}{k_B T_{it}} \right)^{\sqrt{0.25}} \right) \exp\left(\frac{-eU_c}{k_B T_s} \right) \right], \tag{11b}$$

$$\xi_{it}^{pgw} = n_{it}(z) \begin{pmatrix} l_{pgw} t_{pgw} \\ + t_{pgw} h_{pgw} \\ + h_{pgw} l_{pgw} \end{pmatrix} \left(\frac{k_B T_{it}}{2\pi^2 m_{it}} \right)^{\sqrt{0.25}} \times \begin{pmatrix} \frac{2}{\sqrt{\pi}} \left(\frac{eV_{pg}}{k_B T_{it}} \right) \\ + \exp\left(\frac{eV_{pg}}{k_B T_{it}} \right) erfc \left(\frac{eV_{pg}}{k_B T_{it}} \right) \end{pmatrix} \times \exp\left(\frac{-eU_c}{k_B T_s} \right), \tag{11c}$$

D_w and L_w are the average height and diameter of the wth CNT in CNT array. l_{pgw},t_{pgw} and h_{pgw} are the length, thickness and height of PGL on wth catalyst nanoparticle. $\psi_v = \frac{2e^2}{D_w 4\pi\varepsilon_\circ}$, V_{cur} and V_{pg} denotes potential at tip, curved surface of wth CNT (refer equation (A1) of the appendix) and at surface of PGL on wth catalyst nanoparticle [56].

ξ_e^{tipw}, ξ_e^{curw}, ξ_e^{pgw} represents the electron collection current at wth - CNT tip, curved surface and PGL and are defined as,

$$\xi_e^{tipw} = \left[n(z) \pi \frac{D_w^2}{4} \sqrt{\left(\frac{8k_B T_e}{\pi m_e} \right)} \exp\left[\frac{Z_{CNT} \psi_V}{k_B T_e} + \frac{eU_c}{k_B T_s} \right] \right], \tag{11d}$$

$$\xi_e^{curw} = \left[n(z) \frac{D_w}{2} L_w \sqrt{\left(\frac{2\pi k_B T_e}{m_e} \right)} \exp\left[\frac{eV_{cur}}{k_B T_e} + \frac{eU_c}{k_B T_s} \right] \right], \tag{11e}$$

$$\xi_e^{pgw} = n(z)(l_{pgw} t_{pgw} + t_{pgw} h_{pgw} + h_{pgw} l_{pgw}) \left(\frac{k_B T_e}{2\pi^2 m_e} \right)^{\sqrt{0.25}} \left(\exp\left(\frac{eV_{pg}}{k_B T_e} + \frac{eU_c}{k_B T_s} \right) \right), \tag{11f}$$

In equation (11), on right-hand side, the first, second, and third term outlines the accreting positive charges of type j, k, l on CNT array and PGL with time whereas the last term depicts a decrease in charge due to electron accumulation with time.

2.5. Kinetics of electrons in plasma

The time evolution of electrons in the bulk plasma is attributed to ionization of neutral atoms, ion-electron recombination, electron collection current on the hybrid surface, and electron loss to the chamber wall. It is computed using the following equation,

$$\dot{n} = \sum_{t}^{j,k,l} \delta_t n_t - \sum_{t}^{j,k,l} \alpha_t n n_{it} - \gamma_e \Lambda_{hyb} \sum_{w=1}^{r} \begin{pmatrix} \xi_e^{tipw} + \xi_e^{curw} \\ + \xi_e^{pgw} \end{pmatrix} - K_{wall}^e n, \tag{12}$$

where, δ_t is the ionization coefficient of neutral atoms. α_t is the recombination coefficient of electrons and ions, and is expressed as $\alpha_t \left(= \alpha_{to} \left(\frac{300}{T_e} \right)^q \right)$cm^{3}sec^{-1} with q (= −1.2) constant, α_{to} =1.12 × 10^{-7} cm^{3}sec^{-1}. γ_e(≈1) is the sticking coefficient of electrons, Λ_{hyb} represents hybrid number density. $K_{wall}^e n = \left(\frac{\gamma_e \nu_{the} Sn}{4V} \right)$ is the loss of electrons per unit time per unit volume on the chamber wall, ν_{the} is the thermal velocity of electrons (≈5.13 × 10^7 cm sec^{-1}) S is the chamber surface area with inner diameter 32 cm and length 23 cm [57], V denotes the chamber volume (≈1.05 × 10^{4}cm^3).

The first term on the right-hand side of equation (12) represents the increase in electron density over time as a result of ionization of neutral atoms in plasma. The second and third terms represent the electron density timely decay induced by ion-electron recombination and collection current on the CNT-PGL Hybrid The final term represents the electron density loss on the chamber wall per unit time. For electron impact reactions [cf. 58, 59].

Table 5. Description of the terms in equations (13a)–(13c).

Terms	Description
$A_t = \delta_t n_t$	The rate of gain of ions of type t (t corresponds to type j, k, l) due to neutral atom ionization in plasma
$B_t = \alpha_t n n_{it}$	The rate of loss of ions of type t (t corresponds to type j, k, l) due to ion-electron recombination.
$C_t = \Lambda_{hyb} \sum_{w=1}^{r} \begin{pmatrix} \xi_{it}^{tipw} + \xi_{it}^{curw} \\ + \xi_{it}^{pgw} \end{pmatrix}$	The rate of loss of ion of type t (t corresponds to type j, k, l) due to the ion collection current at the hybrid surface.
$D_t = K_{wall}^{it} n_{it} = \left(\frac{\gamma_{it} \nu_{thit} S n_{it}}{4V} \right)$	The rate of loss of ions of type t (t corresponds to type j, k, l) to chamber wall
$E_j = \sum_i P_{ij} n_j n_{il}$	The rate of gain of ions of type j due to reaction between neutral of type j and ion of type l , as listed in table 6
$F_j = \sum_i P_{ij} n_j n_{ik}$	The rate of gain of ions of type j due to reaction between neutral of type j and ion of type k, as listed in table 6.
$G_j = \sum_i P_{il} n_l n_{ij}$	The rate of loss of ions of type l due to reaction between neutral of type l and ion of type j, as listed in table 6.
$E_k = F_{th}$	The rate of gain of ions of type k due to thermal dehydrogenation
$F_k = \sum_i P_{ij} n_j n_{ik}$	The rate of loss of ions of type j due to reaction between neutral of type j and ion of type k, as listed in table 6.
$G_k = \sum_i P_{ik} n_k n_{ik}$	The rate of gain of ions of type k due to reaction between neutral of type k and ion of type k, as listed in table 6.
$E_l = \sum_i P_{il} n_l n_{ij}$	The rate of gain of ions of type l due to reaction between neutral of type l and ion of type j, as listed in table 6.
$F_l = \sum_i P_{il} n_l n_{il}$	The rate of gain of ions of type l due to reaction between neutral of type l and ion of type l, as listed in table 6.
$G_l = \sum_i P_{ij} n_j n_{il}$	The rate of loss of ions of type j due to reaction between neutral of type j and ion of type l, as listed in table 6.
P_{it}	Rate constant for ion-neutral reactions, as listed in table 6.
ν_{thit}	Thermal velocity of ions of type t (t corresponds to type j, k, l) [59]

2.6. Kinetics of positive ions in plasma

The following equations describe the time evolution of positive ions in the bulk plasma due to, the ionization of neutral atoms, electron–ion recombination, ion collection current at the hybrid's surface, loss of ions to the chamber wall, thermal dehydrogenation, and ion-neutral reactions.

$$n_{ij}^{\cdot} = A_j - B_j - C_j - D_j + E_j + F_j - G_j, \tag{13a}$$

$$n_{ik}^{\cdot} = A_k - B_k - C_k - D_k + E_k - F_k + G_k, \tag{13b}$$

$$n_{il}^{\cdot} = A_l - B_l - C_l - D_l + E_l + F_l - G_l, \tag{13c}$$

Description of the terms in equations (13*a*)–(13*c*) is provided in table 5.

2.7. Kinetics of neutrals in plasma

The following equations describe the time evolution of neutrals in bulk plasma due to, electron–ion recombination, neutralization of ions at the hybrid's surface, ionization of neutral atoms followed by neutrals loss to the chamber wall, their accretion, adsorption, desorption over the hybrid surface, and numerous ion-neutral processes.

$$n_j^{\cdot} = H_j + I_j - J_j - K_j - L_j - M_j + N_j - O_j - P_j + Q_j, \tag{14a}$$

$$n_k^{\cdot} = H_k + I_k - J_k - K_k - L_k - M_k + N_k + O_k - P_k, \tag{14b}$$

$$n_l^{\cdot} = H_l + I_l - J_l - K_l - L_l - M_l + N_l - O_l - P_l + Q_l, \tag{14c}$$

Description of the terms in equations (14*a*)–(14*c*) is provided in table 7.

2.8. Generation of hydrogen and carbon species over catalysts surface

The interaction of neutrals and ions of type j, k, and l with catalyst nanoparticles results in the formation of hydrogen radicals (H_a) and carbon species (C_a) on the catalyst surface, as explained in the supplementary information.

See the supplementary information for specifics on the generation of hydrogen and carbon species on the surface of the catalyst.

Table 6. Ion-neutral reactions between various species considered in the present plasma model [59, 60].

Reactions	Rate Constant (P_{it}) (cm^3/sec)
$H^+ + CH_4 \rightarrow CH_3^+ + H_2$	2.30×10^{-9} [60]
$H^+ + CH_3 \rightarrow CH_3^+ + H$	3.40×10^{-9} [60]
$H^+ + NH_3 \rightarrow NH_3^+ + H$	5.20×10^{-9} [60]
$H^+ + C_2H_3 \rightarrow C_2H_3^+ + H$	2.00×10^{-9} [60]
$H_2^+ + H \rightarrow H^+ + H_2$	6.40×10^{-10} [60]
$H_2^+ + H_2 \rightarrow H_3^+ + H$	2.50×10^{-9} [60]
$H_2^+ + C_2H_2 \rightarrow C_2H_2^+ + H_2$	5.30×10^{-9} [60]
$CH_3^+ + C_2H_2 \rightarrow C_3H_3^+ + H_2$	1.15×10^{-9} [60]
$CH_3^+ + NH_3 \rightarrow NH_4^+ + CH_2$	3.04×10^{-10} [60]
$CH_4^+ + H_2 \rightarrow CH_5^+ + H$	3.50×10^{-11} [59]
$CH_4^+ + CH_4 \rightarrow CH_5^+ + CH_3$	1.50×10^{-9} [59]
$CH_4^+ + C_2H_2 \rightarrow C_2H_2^+ + CH_4$	2.72×10^{-9} [59]
$CH_4^+ + C_2H_2 \rightarrow C_2H_3^+ + CH_3$	2.45×10^{-9} [60]
$CH_4^+ + NH_3 \rightarrow NH_3^+ + CH_4$	2.70×10^{-9} [60]
$CH_4^+ + HCN \rightarrow HCNH^+ + CH_3$	1.20×10^{-9} [60]
$C_2H^+ + H_2 \rightarrow C_2H_2^+ + H$	1.25×10^{-9} [60]
$C_2H^+ + CH_4 \rightarrow C_2H_2^+ + CH_3$	3.70×10^{-10} [60]
$C_2H^+ + C_2H_2 \rightarrow C_4H_2^+ + H$	1.85×10^{-9} [60]
$C_2H^+ + HCN \rightarrow C_2H_2^+ + CN$	2.70×10^{-9} [60]
$C_2H_2^+ + H_2 \rightarrow C_2H_3^+ + H$	1.00×10^{-11} [60]
$C_2H_2^+ + CH_4 \rightarrow C_3H_4^+ + H_2$	6.25×10^{-10} [59]
$C_2H_2^+ + C_2H_2 \rightarrow C_4H_2^+ + H_2$	4.90×10^{-10} [60]
$C_2H_2^+ + NH_3 \rightarrow NH_3^+ + C_2H_2$	2.14×10^{-9} [60]
$C_2H_3^+ + HCN \rightarrow HCNH^+ + C_2H_2$	2.90×10^{-9} [60]
$C_2H_3^+ + NH_3 \rightarrow NH_4^+ + C_2H_2$	2.50×10^{-9} [60]
$H^+ + C_4H_2 \rightarrow C_4H_2^+ + H$	2.06×10^{-9} [60]
$C_3H_2^+ + C_3H_4 \rightarrow C_4H_3^+ + C_2H_3$	1.96×10^{-10} [60]
$HCN^+ + CH_4 \rightarrow C_2H_3^+ + NH_2$	2.60×10^{-10} [60]
$NH_3^+ + NH_3 \rightarrow NH_4^+ + NH_2$	2.10×10^{-9} [60]
$N_2^+ + NH_3 \rightarrow NH_3^+ + N_2$	1.94×10^{-9} [60]
$N_2^+ + CH_4 \rightarrow CH_3^+ + N_2 + H$	9.30×10^{-10} [60]
$CN^+ + NH_2 \rightarrow NH_2^+ + CN$	9.10×10^{-10} [60]
$CN^+ + NH_3 \rightarrow HCN^+ + NH_2$	2.00×10^{-9} [59]
$CN^+ + HCN \rightarrow HCN^+ + CN$	2.70×10^{-9} [60]
$HCN^+ + C_2H_2 \rightarrow C_2H_2^+ + HCN$	1.50×10^{-9} [60]
$HCN^+ + NH_3 \rightarrow NH_3^+ + HCN$	1.68×10^{-9} [60]

2.9. Growth rate equation of CNT array

The average growth rate of wth CNT in the CNT array is assessed by the change in the average volume of wth CNT, represented by the left-hand side of equation (15). The outer diameter of the wth CNT is comparable to the diameter of the w^{th} catalyst nanoparticle (D_{Ni}).

The right-hand side of equation (15) evaluates the growth processes of wth CNT in the CNT array. The carbon monomers and carbon clusters that are generated on the catalyst surface because of several surface processes - such as hydrocarbon adsorption, ion-induced and thermal dissociation of adsorbed hydrocarbons, decomposition of hydrocarbon ions, hydrocarbon and hydrogen ions interaction, diffuse over the surface and into the bulk of the wth Ni catalyst nanoparticle with different energy barriers. The temperature of the catalyst nanoparticles has a considerable impact on these diffusion processes. At lower catalyst temperatures, the surface diffusion process dominates CNT array growth, whereas at higher catalyst temperatures, CNT array growth is primarily owing to the bulk diffusion process, and at intermediate temperatures, both processes contribute. As a result, in this paper, we have studied both the diffusion process in CNT array growth because the temperature of the catalyst nanoparticle is assumed constant.

After this, the carbon atoms in wth Ni catalyst bulk precipitate out with an energy barrier for incorporation along the CNT-Ni interface and result in the formation of a graphitic cylindrical tube. The highly active hydrogen species contribute to array expansion by eliminating amorphous carbon from the wth catalyst surface during array formation. The expansion also incorporates the accumulation of neutrals of type j and l throughout the surface of wth CNT [55, 62]. The accumulated neutrals of type l in the CNTs lattice significantly affect the growth of CNTs.

Table 7. Description of the terms in equations (14a)–(14c).

Terms	Description
$H_t = \alpha_t n n_{it}$	The rate of gain of neutrals of type t (t corresponds to type j, k, l) due to electron- ion combination.
$I_t = \Lambda_{hyb}(1-\gamma_{it})\sum_{w=1}^{r}\begin{pmatrix}\xi_{it}^{tipw}+\xi_{it}^{curw}\\+\xi_{it}^{pgrw}\end{pmatrix}$	The rate of gain of neutrals of type t (t corresponds to type j, k, l) due to ion neutralization on hybrid surface.
$J_j = K_{wall}^{t} n_t = \left(\frac{\gamma_t D_t}{\ell}\right)$	The rate of loss of neutrals of type t to chamber wall. (t corresponds to type j, k, l) D_t= Diffusion coefficient of neutrals within plasma, ℓ= Effective diffusion length of cylindrical chamber [34, 61].
$K_t = \delta_t n_t$	The rate of loss of neutrals of type t due to ionization of neutral atoms (t corresponds to type j, k, l).
$L_t = \Lambda_{hyb}(\gamma_t)\sum_{w=1}^{r}\begin{pmatrix}j_{it}^{tipw}+j_{it}^{curw}\\+j_{it}^{pgw}\end{pmatrix}$	The rate of loss of neutrals of type t due to their accretion at hybrid's surface (t corresponds to type j, k, l).
$M_t = F_{adt} = \frac{p_t}{(2\pi m_{it} k_B T_{it})^{\sqrt{0.25}}} \times \frac{n_t}{F_t}$	The rate of loss of neutrals of type t due to their adsorption at hybrid's surface. p_t = partial pressure of adsorbing species (t corresponds to type j, k, l).
$N_t = F_{dest} = n_t f \exp\left(-\frac{U_{ads}}{k_B T_{it}}\right)$	The rate of gain of neutrals of type j due to their desorption from hybrid's surface with adsorption energy $U_{ads;}$ (t corresponds to type j, k, l).
$O_j = \sum_i P_{ij} n_j n_{il}$	The loss of neutrals of type j due to reaction between neutral of type j and ion of type l, as listed in table 6.
$P_j = \sum_i P_{ij} n_j n_{ik}$	The loss of neutrals of type j due to reaction between neutral of type j and ion of type k, as listed in table 6.
$Q_j = \sum_i P_{il} n_l n_{ij}$	The gain of neutrals of type l due to reaction between neutral of type l and ion of type j, as listed in table 6.
$Q_k = \sum_i P_{ij} n_j n_{ik}$	The gain of neutrals of type j due to reaction between neutral of type j and ion of type k, as listed in table 6.
$P_k = \sum_i P_{ik} n_k n_{ik}$	The loss of neutrals of type k due to reaction between neutral of type k and ion of type k, as listed in table 6.
$O_l = \sum_i P_{il} n_l n_{ij}$	The loss of neutrals of type l due to reaction between neutral of type l and ion of type j, as listed in table 6.
$P_l = \sum_i P_{il} n_l n_{il}$	The loss of neutrals of type l due to reaction between neutral of type l and ion of type l, as listed in table 6.
$Q_l = \sum_i P_{ij} n_j n_{il}$	The gain of neutrals of type j due to reaction between neutral of type j and ion of type l, as listed in table 6.

$$\frac{\pi}{4}\frac{d(D_{Niw}^2 - D_w^2)L_w}{dt} = \frac{1}{r}$$
$$\sum_{w=1}^{r}\left[\begin{pmatrix}\frac{D_a D_{Niw}}{f}\exp\left(\frac{-U_{voc}}{k_B T_s}\right)\\+\frac{D_b}{f}\exp\left(\frac{-U_{arc}}{k_B T_s}\right)\\+\frac{D_c}{f}\exp\left(\frac{-U_{arcl}}{k_B T_s}\right)\\\frac{A_s D_{Ni} w}{f}\exp\left(\frac{-\delta U_{INc}}{k_B T_s}\right)\end{pmatrix}\times\left(C_a \times \xi_{ij} \times \left(\frac{M}{\rho}\right)_w\right)\right.$$
$$\left.+\left(\left[\pi D_{Niw} L_w \times H_a \times \xi_{ik} \times \left(\frac{M_{Ni}}{\rho_{Ni}}\right)_w\right] + \left[(\Upsilon_j \xi_j + \Upsilon_l \xi_l) \times \left(\frac{M}{\rho}\right)_w\right]\right)\right], \quad (15)$$

The details of growth processes integrated in equation (15) are in given table 8.

2.10. Growth rate equation of PGL

When the chamber is cooled, carbon atoms begin to diffuse out from the bulk of the wth catalyst nanoparticles and onto its surface, forming graphene nuclei. These nuclei combine and bridge the spaces between nanoparticles to produce PGLs. Subsequently, the adsorbed carbon is then etched over the catalyst surface, resulting in clean graphene layers.

The growth of PGL is computed in terms of areal (thickness (t_{pg}), and height (h_{pg})) change with time, as expressed by the left-hand side of equation (16).

Table 8. Details of growth processes and parameters integrated in equation (15).

Functions/Parameters	Description
$\frac{D_a D_{Niw}}{f}\exp\left(\frac{-U_{voc}}{k_B T_s}\right)$	Bulk diffusion of carbon monomer into wth catalyst nanoparticle.
$\frac{D_b}{f}\exp\left(\frac{-U_{arc}}{k_B T_s}\right)$	Diffusion of carbon monomer over surface of wth catalyst nanoparticle.
$\frac{D_c}{f}\exp\left(\frac{-U_{arcl}}{k_B T_s}\right)$	Diffusion of carbon cluster over surface of wth catalyst nanoparticle.
$\frac{A_s D_{Niw}}{f}\exp\left(\frac{-\delta U_{INc}}{k_B T_s}\right)$	Incorporation of carbon monomer into wth CNT through wth CNT - Ni confluence.
$\left[\pi D_{Niw} L_w \times H_a \times \xi_{ik} \times \left(\frac{M_{Ni}}{\rho_{Ni}}\right)_w\right]$	Etching of amorphous carbon over the wth catalyst nanoparticle surface.
$\left[(\Upsilon_j \xi_j + \Upsilon_l \xi_l) \times \left(\frac{M}{\rho}\right)_w\right]$	Accumulation of neutrals of type j and l throughout the surface of wth CNT
$U_{voc} = (1.6\ eV)$	Energy barriers for bulk diffusion in catalyst nanoparticle [62]
$U_{arc} = (0.3\ eV)$	Energy barriers for surface diffusion of carbon monomer over catalyst surface [62].
$U_{arcl} = (0.48\ eV)$	Energy barriers for surface diffusion of carbon clusters over catalyst surface [63].
$U_{Inc} = (0.4\ eV)$	Energy barrier for incorporation along CNT- Ni interface [62]
D_a	Bulk diffusion coefficients [62]
D_b, D_c	Surface diffusion coefficients [62]
A_s	Incorporation speed of carbon monomer into wth CNT
M (≈12 g)	Mass of wth CNT
ρ	Density of wth CNT
Y_t(≈1)	Sticking coefficient of neutrals of type j and l [62]

$$l_{pg}\frac{d(h_{pg}\times t_{pg})}{dt} = \sum_{w=1}^{r}\left[\left(D_a\frac{(C_1-C_0)}{\sqrt{\pi D_a t}}\pi D_{Niw}^2 + \frac{(C_1-C_0)}{\sqrt{\pi D_a t}}D_b\pi D_{Niw}^2 + y_i I_K n_{it}\vartheta_0\theta_c l_{pgw} h_{pgw}\right)\frac{M_{wpg}}{\rho_{wpg}}\right], \quad (16)$$

where, M_{wpg} (≈12 g) and ρ_{wpg} are the mass and density of graphene layer on wth catalyst nanoparticle. y_i and I_K are the etching yield and rate coefficient for ions incident on the catalyst nanoparticle [64].

The right-hand side of equation (16) includes the growth processes involved in PGL growth. The first phrase refers to the outward diffusion of carbon atoms from the bulk of the wth Ni catalyst nanoparticle, followed by diffusion over the catalyst surface to generate graphene layers with activation energy barriers [65]. The final part of the equation represents the etching of adsorbed carbon atoms over the catalyst surface, promoting the formation of clean graphene.

2.11. Field enhancement factor

Understanding the field emission characteristics requires the study of the field enhancement factor. Equations (17) and (18) show the approximated field enhancement factor of the CNT array and CNT-PGL hybrid.

$$\beta_a \approx \beta\left\{1 - \exp\left(-2.3172\frac{s}{L_w}\right)\right\}, \quad (17)$$

$$\beta_{hyb} \approx \beta_a \times \beta_g \times \exp(-\phi'), \quad (18)$$

where β_a and $\beta_{hy\,b}$ denotes the field enhancement factor of CNT array and CNT-PGL hybrid. Exp $(-\phi')$ refers to the exponential fall in field enhancement factor owing to a decrease in electric field penetration due to graphene coverage over the CNT array, with $\phi = \sigma'\sqrt{\mu}/\sqrt{\varepsilon}$. σ', μ, ε, β_g are the conductivity, electric dipole moment, dielectric constant, and field enhancement factor of PGL. s is the spacing among CNTs [32, 66–71].

3. Results and discussion

In this paper, an analytical model is used to describe the growth of the CNT-PGL hybrid under appropriate conditions i.e., C_2H_2 flow rate, catalyst film thickness, substrate temperature, and cooling rate. The analytical model incorporates first-order simultaneous differential equations for plasma sheath, plasma species densities, carbon and hydrogen species surface concentration, growth rate of CNT array and PGL as well as field enhancement factor of CNT and CNT-PGL hybrid. The equations are solved via Mathematica software. In computations, we use substrate area to be 500 × 500 μm^2 covered with Ni film, with average centre-to-centre

Table 9. Initial number density of various plasma species considered in present model [47, 65, 72, 73].

Species	Mass (1 amu = 1.66×10^{-27} kg)	Number density (cm^{-3})
Electron	0.0005486 amu	7×10^{10}
C_2H_2	26 amu	7×10^{13}
CH_4	16 amu	2×10^{14}
CH_3	15 amu	1×10^{13}
C_2H_3	27 amu	7×10^{13}
C_3H_4	40 amu	3×10^{13}
C_4H_2	50 amu	1×10^{13}
H_2	2 amu	2×10^{15}
H	1 amu	3×10^{13}
H_3	3 amu	5×10^{15}
NH_3	17 amu	5×10^{13}
NH_4	18 amu	2×10^{11}
HCN	27 amu	9×10^{11}
N_2	28 amu	4×10^{9}
$C_2H_2^+$	26 amu	7×10^{8}
CH_3^+	15 amu	2×10^{8}
$C_2H_3^+$	27 amu	2×10^{7}
$C_4H_2^+$	50 amu	7×10^{7}
H_2^+	2 amu	1×10^{8}
H^+	1 amu	1×10^{8}
H_3^+	3 amu	1×10^{8}
NH_3^+	17 amu	9×10^{8}
NH_4^+	18 amu	1×10^{9}
HCN^+	27 amu	1×10^{8}
CN^+	26 amu	1×10^{8}

Table 10. Parameters used in present model [25, 36].

Parameters	Value
Electron temperature (T_{eo})	2.2 eV
Neutral temperature (T_{no})	0.17 eV
Ion temperature (T_{io})	0.15 eV
Plasma Power	120 W
Substrate Potential (U_c)	−500 V
Substrate temperature (T_s)	750 °C
sGas pressure while CNT array growth	7 Torr
Gas pressure while graphene growth	22 Torr
Initial bulk concentration in Ni catalyst nanoparticle (C_1)	0
NH_3 gas flow rate	200 sccm
N_2 flow rate	10,000 sccm

spacing (s) of 1 μm among the catalyst nanoparticles formed. Tables 9 and 10 list relevant initial conditions and parameters that were determined experimentally. They are utilized to solve equations presented in section (2) at t = 0.

In the present analysis, the obtained potential distribution inside the plasma sheath, ion, and neutral density are used to compute the average diameter (D_{Ni}) of the catalyst nanoparticle together with hydrogen and carbon flux over the catalyst surface. The obtained carbon and hydrogen flux are used to calculate the growth rates of planar graphene and CNT arrays. Then we start the next time by solving the differential equations again, with new boundary conditions derived from the dimensions of the recently acquired hybrid.

First, we analyze the effect of C_2H_2 flow rate, catalyst film thickness, and substrate temperature on the growth of the CNT array, followed by an analysis of the growth of PGL over the CNT array by cooling the chamber.

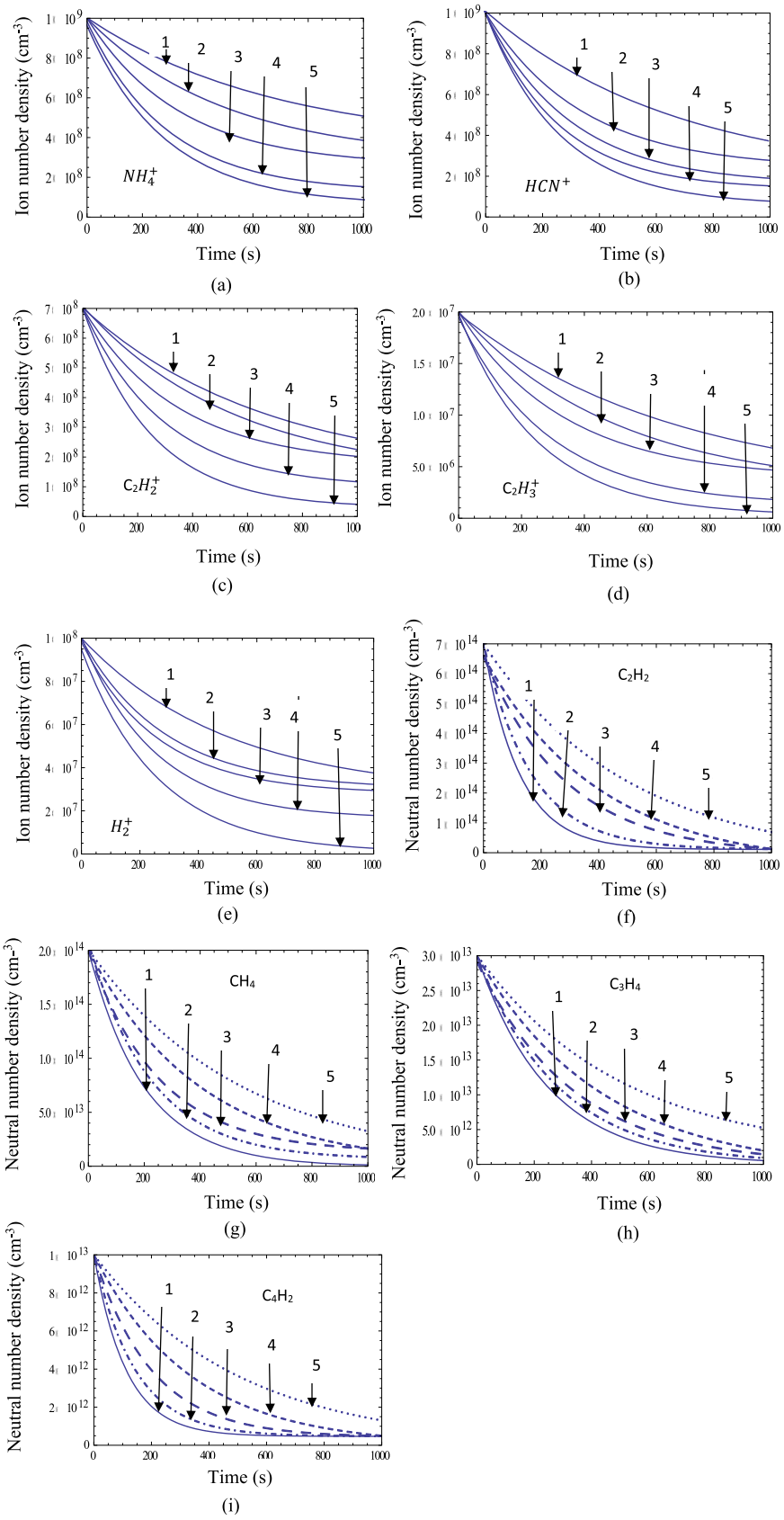


Figure 2. Time varying ion and neutral number density of (a) NH_4^+, (b) HCN^+ (c) $C_2H_2^+$ (d) $C_2H_3^+$ (e) H_2^+ (f) C_2H_2 (g) CH_4 (h) C_3H_4 (i) C_4H_2 varying C_2H_2 flow rate and constant NH_3 flow rate (200 sccm). (1 = 20 sccm), (2 = 40 sccm), (3 = 60 sccm), (4 = 100 sccm) (5 = 150 sccm).

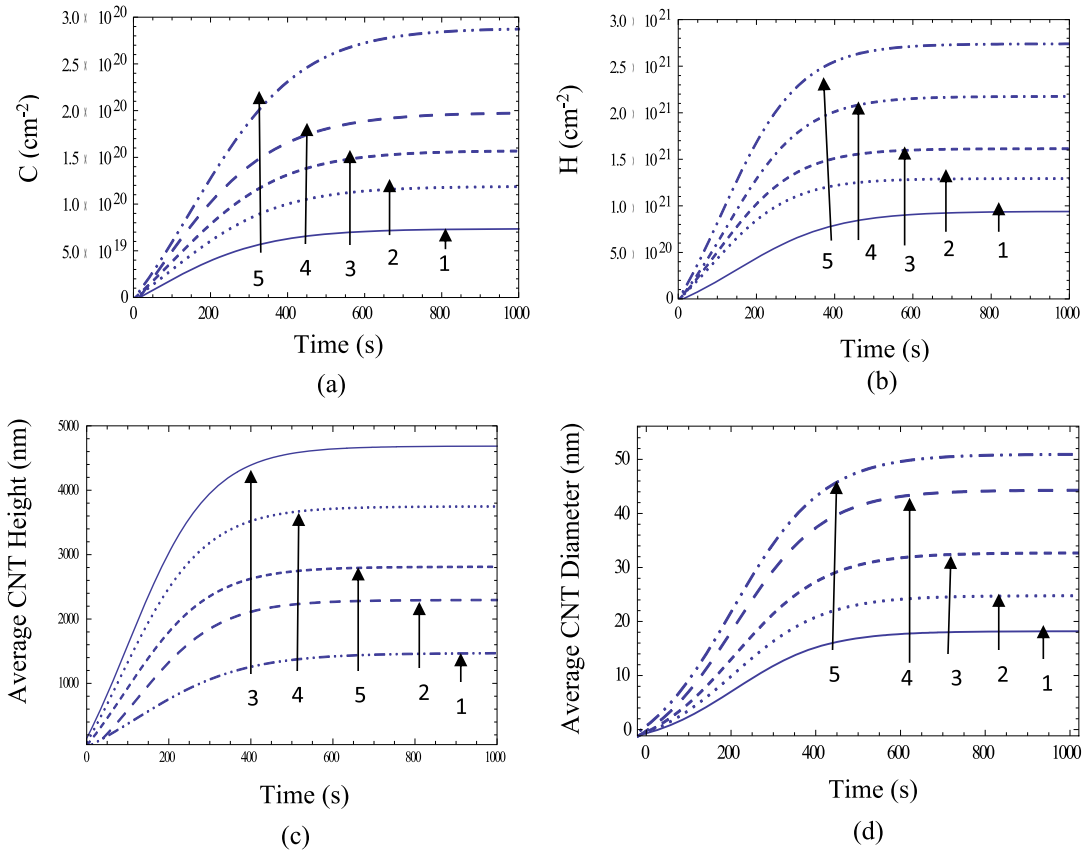


Figure 3. Time varying (a) Surface concentration of Carbon (C_a), (b) Surface concentration of Hydrogen (H_a), (c) Average CNT Height, (d) Average CNT diameter for varying C_2H_2 flow rate and constant NH_3 flow rate (200 sccm). (1 = 20 sccm), (2 = 40 sccm), (3 = 60 sccm), (4 = 100 sccm) (5 = 150 sccm).

3.1. Time variation of CNT array growth with C_2H_2 gas flow rate

Lighter and etchant ions (NH_4^+, HCN^+, $C_2H_2^+$, $C_2H_3^+$, H_2^+) are predominant under low C_2H_2 gas flow rate [cf. figures 2(a)–(e)] assisting in the formation of a clean CNT array [74]. With further increase in C_2H_2 gas flow rate, the neutral number density of type j in plasma bulk increases [cf. figures 2(f)–(i)] leading to their increased adsorption on the catalyst surface. In addition, NH_3 is insufficient to suppress the decomposition of C_2H_2. So low etchant ion density and insufficient NH_3 at high C_2H_2 flow rate results in excess carbon deposition as amorphous carbon, poisoning the catalyst particle, retarding the CNT array growth. Hence, at a low C_2H_2 gas flow rate (i.e. up to 60 sccm) the relative carbon and hydrogen radical surface concentration [cf. figures 3(a), (b)] is competent for an increase in the growth rate of CNT, as carbon species generated over the surface of the catalyst are deposited and amorphous carbon is etched.

Figure 3(c) shows that the average CNT height in the CNT array increases up to a C_2H_2 flow rate of 60 sccm and then it declines with a further rise in C_2H_2 flow at fixed ammonia (NH_3) gas flow rate (200 sccm). Chhowalla *et al* [30] reported an initial increase in CNT growth rate up to 30% C_2H_2 concentration then it dropped further and according to Bell *et al* [31], CNTs are well aligned between 4 and 20% of C_2H_2, with a growth peak at 20%. At 29% C_2H_2, CNTs resemble obelisks while at 38%, amorphous C deposition was detected.

Because of carbon deposition on the catalyst nanoparticle surface, the average diameter of CNTs in an array increases as the flow rate of C_2H_2 increases [cf. figure 3(d)]. The findings are in line with Merkulov *et al* [75], who found a 36 nm tip diameter at a C_2H_2/NH_3 ratio of 0.625 and a 46 nm tip diameter at a ratio of 0.80.

3.2. Variation of CNT array length and diameter with catalyst thickness

Figure 4(a) shows the average CNT height in array with increasing Ni film thickness. As thickness increases (upto 10 nm), catalyst particle size increases, raising average

3.3. Variation of CNT array length and diameter with catalyst thickness

Figure 4(a) shows the average CNT height in an array with increasing Ni film thickness. As thickness increases (up to 10 nm), catalyst particle size increases, raising average CNT height because catalyst particle remain active for dissolution, surface, and bulk diffusion of C- atoms [76]. Above 10 nm thickness, a decrease in average CNT

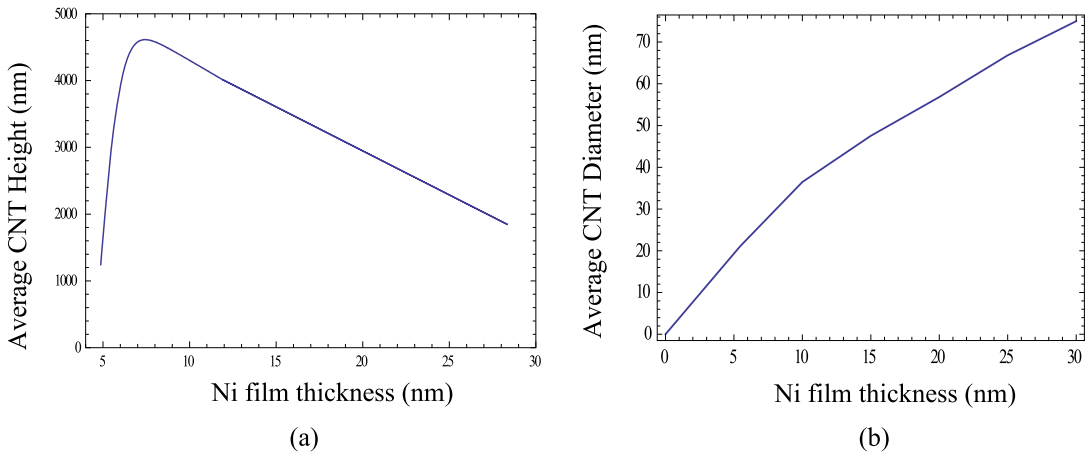


Figure 4. (a) Average CNT Height, (b) Average CNT diameter for varying Ni film thickness.

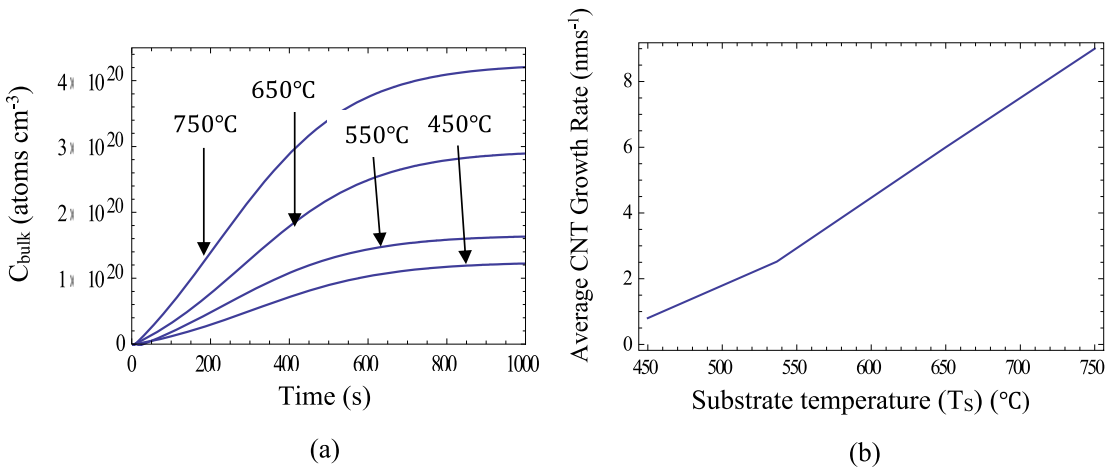


Figure 5. (a) Time varying Carbon atom concentration (C_{bulk}) in Ni nanoparticle. (b) Average CNT growth rate with substrate temperature.

height is observed owing to an increase in diffusion length of carbon inside the catalyst particle, so not enough particles can diffuse out from the catalyst bulk; consequently, shorter nanotubes are obtained. The obtained results comply with the studies of Lee *et al* [76] where an increase of Ni catalyst thickness from 4 to 10 nm, enhances the length and growth rate of CNT by a factor of 0.7. As per Han *et al* [77], the acquired data indicate that, in the case of a 100 nm Ni layer, carbon nanoparticles rather than carbon nanotubes appear in the SEM pictures [cf. figure 2 of Han *et al* [77]]. Ma *et al* [78] conducted an experiment wherein they noticed that the average length of CNTs decreased with increasing Ni film thickness up to 20 nm [cf. figure 4 of Ma *et al* [78]].

Additionally, the size of the Ni nanoparticles, which grow with Ni film thickness, controls the average diameter of CNT [cf. figure 4(b)]. This comes after Han *et al*'s study [77], in which the mean diameter of CNTs increased from 32 to 113 nm when the thickness of the Ni film increased from 10 to 100 nm.

3.4. Variation of growth rate of CNT array with substrate temperature

With increasing substrate temperature, the concentration of carbon atoms (C_{bulk}) inside Ni catalyst nanoparticles increases [cf. figure 5(a)]. This can be attributed to an increase in ion-driven hydrocarbon and ammonia species dissociation and decomposition over Ni catalyst nanoparticle. As a result, C_a and H_a increases across Ni nanoparticle surface. Increases in H_a cause etching, which removes excess carbon from the Ni surface. This opens up the catalyst surface area, allowing a larger intake of carbon and a rise in carbon atom solubility [79]

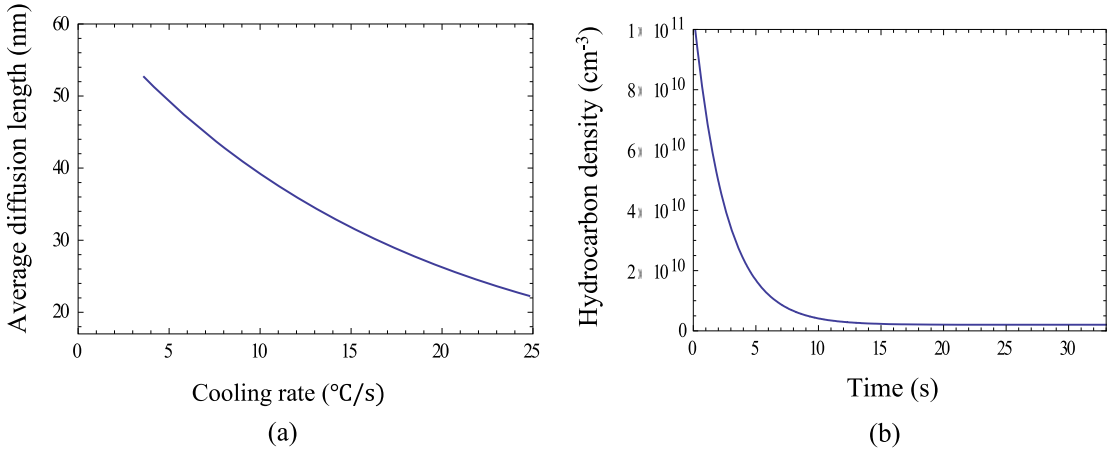


Figure 6. (a) Average diffusion length with varying cooling rates (b) Time varying hydrocarbon density.

and diffusivity. Studies by Bleu *et al* [80], Shurman *et al* [27], have shown that the concentration of carbon in the Ni bulk increases with substrate temperature.

As a result, the average growth rate of CNT in the array rises [cf. figure 5(b)]. This is in line with research done by Chhowalla *et al* who found that CNT height increased between 520 °C and 700 °C [30].

3.5. Variation of: diffusion length with cooling rate, hydrocarbon density with time in N_2 environment to obtain PGL

The carbon diffusion length inside Ni nanoparticles is explored for graphene precipitation. This is done in the N_2 environment after stopping C_2H_2/NH_3 flow. At a slow cooling rate (5 °C s^{-1}), the diffusion length is larger than the size of Ni nanoparticles, which leads to Ni completely carburizing and inhibits the formation of graphene [45]. At a rapid cooling rate (25 °C s^{-1}), a shorter diffusion length is obtained. Therefore, carbon atoms near the Ni surface precipitate more quickly than they diffuse in the Ni bulk, resulting in the formation of amorphous carbon [81]. Figure 6(a) illustrates the effective average diffusion length of carbon into Ni nanoparticle bulk at slow, moderate (15 °C s^{-1}), and fast cooling rates. The calculated diffusion lengths agree with the results of Gkouzou *et al* [45]. According to Yu *et al* [26] and Reina *et al* [81], the average diffusion length decreases with a rise in cooling rates.

Starting N_2 flow and stopping C_2H_2/NH_3 flow after the formation of the CNT array, causes a gradual decrease for species remaining in the chamber [cf. figure 6(b)]. This is explained by the fact that N_2 helps to eliminate volatile byproducts such as HCN, C_2N_2, CN, and volatile hydrocarbons as well as any remaining unreacted precursors [82, 83]. This prevents the catalyst from contamination during the precipitation process. As a result, precipitation processes dominate surface reaction processes in the formation of PGL.

With time as the temperature of the chamber decreases at a rate of 15 °C s^{-1}, the carbon atoms start precipitating out from the Ni nanoparticle catalyst bulk. The excess carbon atoms that precipitate out to form graphene layers are proportionate to the difference in concentration of the saturated (C_1) and supersaturated (C_0) solution of carbon atoms inside the Ni catalyst [45]. More carbon precipitates out when the temperature drops because the concentration difference between saturated and supersaturated solutions increases and acts as a constant driving force [44, 45].

Consequently, the synthesis of graphene is guided by the outward and surface diffusion of carbon atoms from and on the Ni surface [46, 64, 84]. Therefore, we can depict that the number of graphene layers; hence the thickness of graphene increases with time on cooling [27, 49], [cf. figures 7(a), (b)]. The experimental work of Choi *et al* [25] regarding number of graphene layers (4–7 layers), is similar to the obtained value (≈5 layers). Reina *et al* [49] have obtained thickness within the range of 1–5 nm corresponding to approximately 1–12 graphene layers. Gupta *et al* [85] have plotted the thickness of the graphene layer for different values of graphene layer number [cf. figure 1 of Gupta *et al* [85]].

3.6. Variation of graphene layer number with Ni film thickness

At Ni film thicknesses of less than 3 nm, small catalyst nanoparticles form. Carbon atoms rapidly saturate these catalyst nanoparticles by forming a carbon shell around them [76]. The catalyst nanoparticle size rises with the thickness of the Ni film, causing the CNT array to grow due to increased carbon atom dissolution and diffusion [76]. This simultaneously increases the number of carbon atoms available for precipitation, eventually leading to

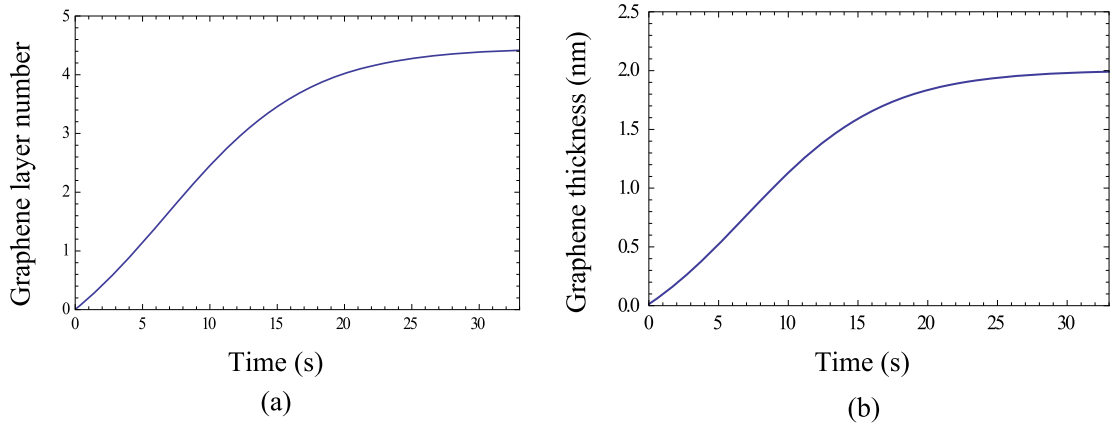


Figure 7. (a) Time varying (a) graphene layer number (b) graphene thickness.

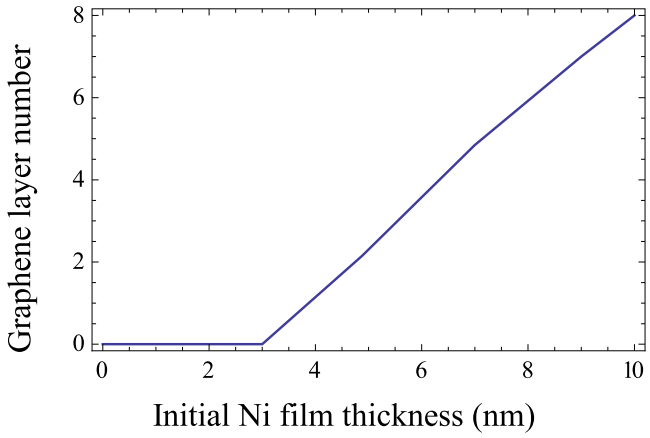


Figure 8. Varying graphene layer number with initial Ni film thickness.

Table 11. Optimized dimensions of CNT array and CNT-PGL hybrid considered.

CNT array	CNT –PGL Hybrid
Average height of CNT (L_w) ≈ 5 μm	Thickness of graphene (t_{pg}) ≈ 2 nm
Average diameter of CNT (D_w) ≈ 32 nm	Average height of CNT ≈ 5 μm
Spacing among CNT ≈ 1 μm	Average diameter of CNT ≈ 32 nm

the formation of graphene layers [46]. Up to a film thickness of 3 nm, figure 8 displays no graphene layer; beyond that, the layer number starts to rise with an increase in film thickness. For a 7 nm thick Ni film, Choi *et al* [25] got 4–7 graphene layers; Peng *et al* [46] and Kim *et al* [86], have shown that the number of graphene layers increases with the thickness of the Ni film.

3.7. Field enhancement factor of CNT array and CNT-PGL hybrid

The procured results from the present study are utilized to determine the field enhancement factor of the CNT-PGL hybrid (β_{hyb}) and to compare it with the field enhancement factor of CNT array (β_a). Table 11 lists the dimensions of the CNT array and CNT-PGL hybrid obtained at a C_2H_2 flow rate of 60 sccm, Ni film thickness of 8 nm, substrate temperature 750 °C, cooling rate 15 °C s^{-1}.

The obtained β_a [cf. figure 9] is consistent with the experimental results of Bonard *et al* [32], who got β_a (≈800) for the following parameters of CNT film: mean height 3 μm, mean tube radius 19 nm, and inter-tube distance 1.02 μm. Furthermore, it is observed that β_a is less than β_{hyb}. Because CNTs are densely packed, the higher coulomb-repelling force of neighboring CNTs results in a significant reduction of β_a. The large β_{hyb} is

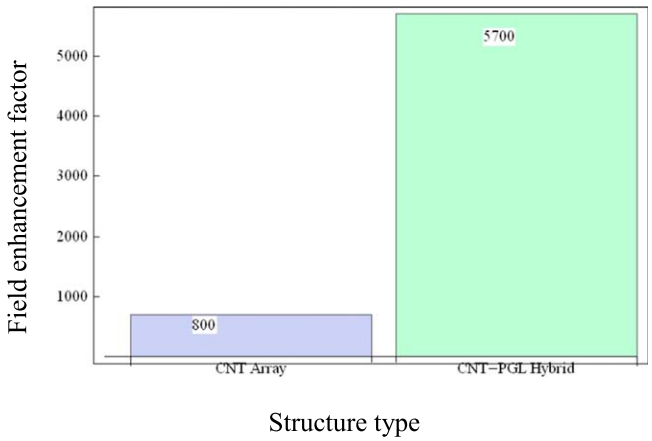


Figure 9. Field enhancement factor of CNT array and CNT-PGL Hybrid.

ascribed to the field emission from (i) the PGL's edges and (ii) the spaces between CNTs since PGL is present on CNT [86, 87]. Therefore, precipitating PGL across a CNT array considerably increases the field enhancement factor, exceeding βa. Obtained β_{hyb} is following Qu *et al* [88] who have obtained β_{hyb} ($\approx$ 5780). Kaur *et al* [33] [cf. table 2 of Kaur *et al* [33]], Hong *et al* [19], Kaushik *et al* [87] [cf. table 3 of Kaushik *et al* [87]] have also reported an escalation in β_{hyb} than β_a.

4. Conclusion

The model can configure and characterize the growth of the CNT-PGL hybrid by adjusting the C_2H_2 flow rate, catalyst film thickness, substrate temperature, and cooling rate. The effect of varying C_2H_2 flow rates on ion and neutral number density, as well as on carbon and hydrogen radical surface concentration is evaluated, to investigate their influence on the growth of the CNT array. It has been noted that as the C_2H_2 flow rate and Ni film thickness increase, the growth rate of the CNT array first increases and subsequently drops. Furthermore, the solubility and diffusivity of carbon atoms inside the Ni catalyst increase with substrate temperature, enhancing the average growth rate of CNTs. Moreover, as chamber temperature decreases at a rate of 15 °C s^{-1} and catalyst film thickness increases the number of graphene layers grows. The procured results are employed to determine the field enhancement factor of densely packed CNT array and CNT-PGL hybrid and it is concluded that precipitated graphene layers on the CNT array top can enhance the field enhancement factor of the densely packed CNTs. Thus gas flow rate, catalyst film thickness, substrate temperature, and cooling rate may all be regulated to control the formation of a CNT-PGL hybrid, which may then be employed in efficient field emission devices.

Acknowledgments

The author (Shruti Sharma) is thankful to Plasma Nano-Simulation Lab, Department of Applied Physics, Delhi Technological University (DTU), India for all the assistance provided during the research work. The author declare that they have no competing interests. No funds, grants or other support was received from external source.

Data availability statement

All the data that support the findings of this study are included within the article, supplementary information and appendix. There is no additional data available beyond what has already been presented.

Ethical statement

Not applicable

Appendix. Potential at curved surface of wth CNT (V_{cur}) [89]

$$V_{cur} = \frac{\rho_q}{4\pi\varepsilon_o}\iiint \frac{1}{\sqrt{(x'^2 + y'^2 + (z - z')^2}} d\tau'$$

$$x' = r'\cos\varphi',\ y' = r'\sin\varphi',\ z = z'$$

$$\frac{\rho_q}{4\pi\varepsilon_o}\int_{-L_w/2}^{L_w/2}\int_0^{2\pi}\int_0^{R_w} \frac{1}{\sqrt{((r'\cos\varphi)'^2 + (r\sin\varphi)'^2 + (z - z')^2}} r'dr'd\varphi'dz'$$

where ρ_q is charg density on curved surface of wth CNT, R_w is the radius of wth CNT.

$$\frac{\rho_q}{4\pi\varepsilon_o}\int_{-L_w/2}^{L_w/2}\int_0^{2\pi}\left[\int_0^{R_w} \frac{r'}{\sqrt{r'^2 + (z - z')^2}} dr'\right] d\varphi'dz'$$

on solving

$$V_{cur} = \frac{\rho_q}{4\pi\varepsilon_o}\left[-2L_w z + \left(z + \frac{L_w}{2}\right)\sqrt{\left(R_w^2 + \left(z + \frac{L_w}{2}\right)^2\right)}\right.$$

$$\left. -\left(z - \frac{L_w}{2}\right)\sqrt{\left(R_w^2 + \left(z - \frac{L_w}{2}\right)^2\right)} + R_w^2 \ln \frac{\sqrt{\left(R_w^2 + \left(z + \frac{L_w}{2}\right)^2\right)} + \left(z + \frac{L_w}{2}\right)}{\sqrt{\left(R_w^2 + \left(z - \frac{L_w}{2}\right)^2\right)} + \left(z - \frac{L_w}{2}\right)}\right] \quad \text{(A1)}$$

ORCID iDs

Suresh C Sharma https://orcid.org/0000-0002-8772-4524

References

[1] Sang M, Shin J, Kim K and Yu K J 2019 *Nanomater* **9** 374
[2] De Volder M F L, Tawfick S H, Baughman R H and Hart A J 2013 *Science* **339** 535
[3] Shim W, Kwon Y, Jeon S Y and Yu W R 2015 *Sci Rep.* **5** 16568
[4] Deng J H, Liu R N, Zhang Y, Zhu W X, Han A L and Cheng G A 2017 *J. Alloys Compd.* **723** 75–83
[5] Gong J, Yang H and Yang P 2015 *Compos B Eng* **75** 250–5
[6] Koh A T T, Chen T, Pan L, Sun Z and Chua D H C 2013 *J. Appl. Phys.* **113** 174909
[7] Parmee R J, Collins C M, Milne W I and Cole M T 2015 *Nano Converg* **2** 1
[8] Ji D, Jang J, Park J H, Kim D, Rim Y S, Hwang D K and Noh Y Y 2020 *J. Inf. Disp.* **22** 1–11
[9] Kumar V 2021 *Trans Electr Electron Mater* **22** 515–8
[10] Fang RP, Chen K, Yin LC, Sun ZH, Li F and Cheng H M 2019 *Adv. Mater.* **31** 1800863
[11] Seman R N A R, Azam M S and Mohamad A A 2017 *Renew. Sustain. Energy Rev.* **75** 644–59
[12] Kumar Y A, Koyyada G, Ramachandran T, Kim J H, Sajid S, Moniruzzaman M, Alzahmi S and Obaidat I M 2023 *Nanomater* **13** 1049
[13] Komatsubara K *et al* 2022 *ACS Appl. Nano Mater.* **5** 1521–32
[14] Ferrier D C and Honeychurch K C 2021 *Biosensors-Basel* **11** 486
[15] Levchenko I, Xu S, Teel G, Mariotti D, Walker M L R and Keidar M 2018 *Nat. Commun.* **9** 879
[16] Singh L A, Walker MLR, Sanborn G P, Turano S P and Ready W J 2013 Operation of spindt-type, carbon nanotube cold cathodes in a Hall effect thruster environment paper presented at the 33rd*Int. Electric Propulsion Conf.* paper IEPC-2013-348
[17] Yap Y K, Menda J, Vanga L K, Kayastha V, Wang J, King L B, Dimovski S and Gogotsi Y 2004 *M.R.S Online Proc Libr* **821** 147–51
[18] Liu J, Zeng B, Wang X, Wang W and Shi H 2013 *Appl. Phys. Lett.* **103** 053105
[19] Hong X, Shi W, Zheng H and Liang D 2019 *Vacuum* **169** 108917
[20] Neyts E C 2012 *J. Vac. Sci. Technol.* B **30** 030803
[21] Marvi Z, Xu S, Foroutan G and Ostrikov K 2015 *Phys. Plasmas* **22** 013504
[22] Deng J H, Cheng G A, Zheng R T, Yu B, Li G Z, Hou X G, Zhao M L and Li D J 2014 *Carbon* **67** 525–33
[23] Wei Y Y, Eres G, Merkulov V I and Lowndes D H 2001 *Appl. Phys. Lett.* **78** 1394
[24] Yesilbag Y O, Yesilbag F N T, Huseyin A, Tuzluca M, Ismail I and Ertugrul M 2021 *J Mater Sci: Mater Electron* **32** 7943–55
[25] Choi J W, Youn S K and Park H G 2013 *J. Nanomater.* **2013** 734686
[26] Yu Q, Lian J, Siriponglert S, Li H, Chen Y P and Pei S S 2008 *Appl. Phys. Lett.* **93** 113103
[27] Al- Shurman K M and Naseem H 2014 CVD Graphene Growth Mechanism on Nickel Thin Films *Excerpt from the Proc. of the 2014 COMSOL Conf. in Boston*
[28] Lee M W, Haniff M A S M, Teh A S, Bein D C S and Chen S K 2015 *J. Exp. Nanosci.* **10** 1232–41
[29] Thapa A, Jungjohann K L, Wang X and Li W 2020 *J. Mater. Sci.* **55** 2101–17
[30] Chhowalla M, Teo K B K, Ducati C, Rupesinghe N L, Amaratunga G A J, Ferrari A C, Roy D, Robertson J and Milne W I 2001 *J. Appl. Phys.* **90** 5308–5317
[31] Bell M S, Teo K B K, Lacerda R G, Milne W I, Hash D B and Meyyappan M 2006 *Pure Appl. Chem.* **78** 1117–25
[32] Bonard J M, Weiss N, Kind H, Stöckli T, Forró L, Kern K and Châtelian A 2001 *Adv. Mater.* **13** 184–8

[33] Kaur G, Pulagara N V, Kumar R and Lahiri I 2020 *Diam. Relat. Mater.* **106** 107847
[34] Mao M and Bogaerts A 2010 *J. Phys. D: Appl. Phys.* **43** 205201
[35] Bell M S, Lacerda R G, Teo K B K, Rupesinghe N L, Amaratunga G A J, Milne W I and Chhowalla M 2004 *Appl Phy Lett* **85** 1137
[36] Gulas M 2007 Synthesis of Carbon Nanotubes by Plasma Enhanced Hot-Filament Catalytic CVD: Study of Gas Phase Chemistry *WDS'07 Proc. of Contributed Papers, Part III* 123–7
[37] Cruden B A, Cassell A M, Hash D B and Meyyappan M 2004 *J. Appl. Phys.* **96** 5284
[38] Mehdipour H, Ostrikov K and Rider A E 2010 *Nanotechnology* **21** 455605
[39] Mehdipour H, Ostrikov K, Rider A E and Han Z 2011 *Plasma Process Polym* **8** 386–400
[40] Gupta R and Sharma S C 2020 *Plasma Chem. Plasma Process.* **40** 1331–50
[41] Cui H, Yang X, Simpson M L, Lowndes D H and Varela M 2004 *Appl. Phys. Lett.* **84** 4077
[42] Wang X D, Vinodgopal K and Dai G P 2019 *Synthesis of Carbon Nanotubes by Catalytic Vapor Deposition* (IntechOpen)
[43] Thapa A, Neupane S, Guo R, Jungjohann K L, Pete D and Li W 2018 *Diam. Relat. Mater.* **90** 144–53
[44] McCarty K F, Feibelman P J, Loginova E and Bartelt N C 2009 *Carbon* **47** 1806–13
[45] Gkouzou A, Janssen G C A M and van Spengen W M 2020 *Sens Actuator A: Phys* **303** 111837
[46] Peng K J, Wu C L, Lin Y H, Liu Y J, Tsai D P, Pai Y H and Lin G R 2013 *J. Mater. Chem.* C **1** 3862–70
[47] Sodha M S, Misra S and Mishra S K 2009 *Phys. Plasmas* **16** 123705
[48] Ostrikov K and Xu S 1807 *Plasma-Aided Nanofabrication From Plasma Sources to Nanoassembly* (Wiley Publishers Since) (https://download.e-bookshelf.de/download/0000/6030/70/L-G-0000603070-0002364804.pdf)
[49] Reina A, Jia X, Ho J, Nezich D, Son H, Bulovic V, Dresselhaus M S and Kong J 2009 *Nano Lett.* **9** 30–5
[50] Liu H and Liu Y 2017 *Phys Sci Rev* **2** 20160107
[51] John and Marica Price College of Engineering, The University of Utah, Department of Material Science and Engineering, The Zang Research Group 2023 Available Online (https://my.eng.utah.edu/~lzang/images/lecture-4.pdf)
[52] Fang W 2012 Bilayer graphene growth by low pressure chemical vapor deposition on copper foil. (https://dspace.mit.edu/bitstream/handle/1721.1/75656/818359166-MIT.pdf? sequence=2&isAllowed=y)
[53] Gupta R, Sharma S C and Sharma R 2017 *Plasma Source Sci. Technol.* **26** 024006
[54] Sodha M S, Misra S, Mishra S K and Srivastava S 2010 *J. Appl. Phys.* **107** 103307
[55] Gupta R, Gupta N and Sharma S C 2018 *Phys. Plasmas* **25** 043504
[56] Trott M 3D Charges and Configurations with sharp edges (September 27,2012) Wolfram Available Online (https://blog.wolfram.com/2012/09/27/3d-charges-and-configurations-with-sharp-edges/)(accessed on 14 April 2024)
[57] Denysenko I B, Xu S, Long J D, Rutkevych P P, Azarenkov N A and Ostrikov K 2004 *J. Appl. Phys.* **95** 2713
[58] Carrasco E, Redondo M J, Tanarro I and Herrero V J 2011 *Phys. Chem. Chem. Phys.* **13** 19561–19572
[59] Alman D A, Ruzic D N and Brooks J N 2000 *Phys. Plasmas* **7** 1421
[60] The UMIST database for astrochemistry 2012 Available online for database of neutral -ion reactions rate coefficients. http://udfa.net/
[61] da Cruz Baptista Dias A M 2012 *Modeling of Low Pressure Plasmas in CH4-H2 Mixtures* (Instituto Superior Tecnico, Universidade de Lisboa)
[62] Gupta R and Sharma S C 2019 *Contrib. Plasma Phys.* **59** 72–85
[63] Li J 2014 *Thesis 'Theoretical Investigations of Catalytic Methane Cracking and Carbon Nanotube Growth'* (UWSpace) Available Online (https://uwspace.uwaterloo.ca/bitstream/handle/10012/8629/N_Li_Jingde.pdf?sequence=1&isAllowed=y) (accessed on 14 April 2024)
[64] Sharma S and Sharma S C 2022 *Plasma Chem. Plasma Process.* **42** 413–33
[65] Liu N, Fu L, Dai B, Yan K, Liu X, Zhao R, Zhang Y and Liu Z 2011 *Nano Lett.* **11** 297–303
[66] Lv S, Li Z, Liao J, Wang G, Li M and Miao W 2015 *Sci Rep.* **5** 15035
[67] Milovzorov D E 2018 Local electric fields in dielectric and semiconductors ed M S Kandelousi *Electric Field* (IntechOpen)
[68] Kim T S, An Y J, Kim K H, Chung W S and Cho Y R 2006 *Met. Mater. Int.* **12** 339–43
[69] Santos E J G and Kaxiras E 2013 *Nano Lett.* **13** 898–902
[70] Murata H, Nakajima Y, Saitoh N, Yoshizawa N, Suemasu T and Toko K 2019 *Sci Rep.* **9** 4068
[71] Liu X, Wang C Z, Hupalo M, Lin H Q, Ho K M and Tringides M C 2013 *Crystals* **3** 79–111
[72] Hash D, Bose D, Govindan T R and Meyyappan M 2003 *J. Appl. Phys.* **93** 6284
[73] Teo K B K *et al* 2004 *Nano Lett.* **4** 921–6
[74] Mao M and Bogaerts A 2010 *J.Phys.D: Appl. Physs.* **43** 315203
[75] Merkulov V I, Hensly D K, Melechko A V, Guillorn M A, Lowndes D H and Simpson M L 2002 *J. Phys. Chem.* B **106** 10570–7
[76] Lee M W, Haniff M A S M, Teh A S, Bien D C S, Chen S K, Talib Z A and Shaari A H 2015 *Adv. Mater. Res.* **1107** 314–9
[77] Han J H, Kim H J, Yang M H, Yang C W, Yoo J B, Park C Y, Song Y H and Nam K S 2001 *Mater. Sci. Eng.* C **16** 65–8
[78] Ma W, Zhou Z, Li G and Li P 2016 *High Temp Mater Proc* **35** 857–63
[79] Baraton L, He Z B, Lee C S, Cojocaru C S, Châtelet M, Maurice J L, Lee Y H and Pribat D 2011 *EPL* **96** 46003
[80] Bleu Y, Barnier V, Christien F, Bourquard F, Loir A S, Garelie F and Donnet C 2019 *Carbon* **155** 410–20
[81] Reina A, Thiele S, Jia X, Bhaviripudi S, Dresselhaus M S, Schaefer J A and Kong J 2009 *Nano Res.* **2** 509–16
[82] Wilson R L 2017 Deposition of ultra-thin metal oxide films for gas sensing applications *Doctoral thesis (Ph.D), UCL* (University College London)
[83] Schlüter M, Hopf C and Jacob W 2008 *New J. Phys.* **10** 053037
[84] Weatherup R S, Bayer B C, Blume R, Baehtz C, Kidambi P R, Fouquet M, Wirth C T, Schlögl R and Hofmann S 2012 *ChemPhysChem* **13** 2544–9
[85] Gupta A, Chen G, Joshi P, Tadigadapa S and Eklund P C 2006 *Nano Lett.* **6** 2667–73
[86] Kim K S, Zhao Y, Jang H, Lee S Y, Kim J M, Kim K S, Ahn J H, Kim P, Choi J Y and Hong B H 2009 *Nature* **457** 706–10
[87] Kaushik V, Shukla A K and Vankar V D 2014 *Appl. Phys.* A **117** 2197–205
[88] Qu K *et al* 2013 IEEE 14th*Int. Vacuum Electronics Conf. (IVEC)* 2013, 1–2
[89] Griffiths D J 2013 *Introduction to Electrodynamics* 4th edn.(Pearson)

Optical and Quantum Electronics (2024) 56:1246
https://doi.org/10.1007/s11082-024-07176-y

As_{38} Se_{62} based segmented clad-graded index photonic crystal fiber for supercontinuum generation covering 3–9.5 μm with moderate peak power

Akash Khamaru[1] · **Ajeet Kumar**[1]

Received: 2 April 2024 / Accepted: 8 June 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
A novel Segmented Clad-Graded Index (SC-GI) photonic crystal fiber with a circular core composed of $As_{38}Se_{62}$ is proposed to realize the supercontinuum spectrum in the entire transparency range of $As_{38}Se_{62}$ (2–10 μm). SC-GI PCF conveys dual Zero Dispersion Wavelengths at around 2.4 μm and 6.4 μm. The devised PCF also offers a very small effective mode area of 2.46 μm^2 and a very high nonlinear coefficient of 12,005 W^{-1} km^{-1} at the pump wavelength of 2.4 μm. Detailed numerical simulations reveal that a short length of SC-GI PCF (~28 mm) is capable of showing a broad SC spectrum ranging from 3 to 9.5 μm (−30 dB level of power) using an input source of moderate peak power (1.8 kW) pumping at 3.4 μm. To the best of the author's knowledge, this is the first time report of SCG in Segmented Cladding Fiber design. The evaluated SC covers the transparency window of the background material and also covers the mid-IR region. This type of high broadening opens the gateway for applications like optical metrology, optical coherence tomography (OCT), food quality control, telecommunications and future applications covering the Mid-IR region.Please confirm the corresponding author is correctly identified and amend if necessary.Yes, Corresponding author is correctly mentioned.

Keywords Zero dispersion wavelength · Nonlinear coefficient · Photonic crystal fiber · Supercontinuum generation

1 Introduction

Over the last fifty years, communication and data transmission have perceived a huge technological advancement. One of the reasons behind such advancement is the invention of optical fibers. Optical fibers work on the principle of total internal reflection which is maintained by refractive index difference between core-cladding interfaces. For increasing the

✉ Ajeet Kumar
ajeetdph@dtu.ac.in

Akash Khamaru
akashkhamaru@gmail.com

[1] Advanced Photonics Simulation Research Laboratory, Department of Applied Physics, Delhi Technological University, Bawana Road, Shahbad Daulatpur, Delhi 110042, India

Published online: 26 June 2024

influence of refractive indices of materials, optical fibers are upgraded and transformed into a new substitute—photonic crystal fibers. PCFs are designed by incorporating different type of holes composed of air or material with low refractive index arranged in ordered manner around the core. Photonic crystal fiber delivers extraordinary properties like extremely large-mode-area, extremely small-mode-area, low confinement loss, high birefringence, tailored dispersion properties and high magnitude of nonlinearity and many more (She et al. 2021; Matsui 2005; Medjouri 2021; Kim 2020; Dussardier 2011). Quality to provide an extremely small mode area for light confinement and flexibility to choose high nonlinear material guarantees the presence of fascinating various nonlinear effects like self-phase modulation (SPM), cross-phase modulation (XPM), four-wave mixing (FWM), stimulated Raman and Brillouin scattering (SBS-SRS) in PCF. These nonlinear effects can play simultaneously to give rise to a wide band of supercontinuum spectra. When any picosecond or femtosecond laser pulse interacts with media containing inherent nonlinear properties results in the generation of broadband SC covering the whole Mid-IR (MIR) region. Such broadband SC mainly wraps MIR regime applicable in various fields like telecommunications, optical coherence tomography (OCT) (Hartl et al. 2001), generation of frequency combs, and optical metrology (Ye 2003). To gather the best results of SC, one need to take care of two aspects in PCF: i) arrangement of the air-holes around the core. These air holes altogether control the dispersion characteristics of fiber and hence help in broadening of spectrum. ii) Material selection for composition of core and cladding. For enhancing nonlinear effects such as SCG, core should be composed of material inherits very high refractive index and high nonlinearity. Nonlinearity of silica is in the range of 10^{-20} m^2W^{-1} which bounds its application only within near-IR range (Wang 2023). Dudley et al. had designed ASMF (Air-Silica Micro-structured Fibers) and simulated at pumping of pulse-width 100 fs at 780 nm resulted SC spectra of 500–1400 nm (Alfano and Shapiro 1970). To exaggerate the SC in Mid-IR or Far-IR regions, silica is begun to be replaced with tellurite (Xu et al. 2018; Zheng 2021; Liao 2009), chalcogenide (Siwach 2018; Saini 2017; Saini 2015) chalcopyrite (Salimullah 2023) etc. In recent times, chalcogenide glass formed PCFs are at boom for creation of SC as they offer transparency range of 2–25 µm and inherits high nonlinearity of range 10^{-17} m^2W^{-1}. Diouf et al. had presented two optical structures (one is PCF and another is rib waveguide) composed of $AsSe_2/As_2Se_5$ glasses offering SC band expanded up to 13 µm for the former structure and 16 µm for the latter (Diouf et al. 2019a, b). Van et al. had proposed a hexagonal core PCF with the composition of As_2Se_3 as a background component delivering SC spanning up to 10 µm with the input of 3.5 µm pumping (Chu et al. 2022). $Ge_{15}Sb_{15}Se_{70}$ fabricated PCF giving the spanned SC of 1–7 µm by Medjouri et al. 2019. By the doping of iodine in $Ge_{10}As_{22}Se_{68}$ chalcogenide glass, Zero dispersion wavelength was shifted from 7 µm to 5.5 µm and SC broadens from 1.2 to 15.2 µm at 12 cm fiber reported by Jiao et al. (2019). Wang et al. simulated the SC spectra (2.3–14.5 µm) with a 17 cm long Te-rich chalk ogenide PCF (Wang et al. 2019). $Ge_{11.5}As_{24}Se_{64.5}$ based PCF was proposed by Vyas et al. which provides SC of 1–15 µm at fiber length of 100 mm with a pulse peak power of 3 kW (Vyas 2016). Among all the Selenium (Se) composed glasses $As_{38}Se_{62}$ show high nonlinearity, deliver a nonlinear refractive index of high magnitude (Khamaru 2023) and offer a range of transparency of 2–10 µm. Moller et al. were the first to propose this particular material for SC formation and gave proven results of 1.7–7.5 µm SC band (Møller et al. 2015) In very recent times, Vyas et al. showed a broadband SC of 2–9 µm in the same material at 4 µm pump wavelength with 200 mm of fiber length (Vyas et al. 2021).

In the present work, we have designed a segmented clad-graded index (SC-GI) circular core PCF to realize the SCG. In the author's knowledge, this is the first time when a PCF

having segmented cladding is being explored for SCG. Our numerical simulations deliver favorable dispersion properties consisting of dual zero dispersion wavelengths (2.4 μm and 6.4 μm). This empowers SC-GI PCF to achieve SCG in a broadband range spanning from 3 to 9.5 μm with low peak power (1.8 kW) and short fiber length (~28 mm), which totally covers the transparency regime of $As_{38}Se_{62}$ glass and almost covers the Mid-IR region and opens the gateway for new applications and advancements over them.

The paper is organized into broad five parts. Section 1 comprises the introduction and brief survey of past works presented on SCG in PCFs. The methodology for the analysis is given in Sect. 2. Section 3 gives the structural explanation of the proposed SC-GI PCF for SCG applications. All the results of the simulation and respective discussions are done in Sect. 4. Finally, conclusions are drawn and encapsulated in Sect. 5.

2 Methodology

For simulating SC-GI PCF, we have utilized the FEM-based full vectorial modelling software package COMSOL Multiphysics. The FEM method has been used for the complete modal analysis of the design. Supercontinuum is generated with the inclusion of both linear and nonlinear characteristics of the fiber which are described in further sub-sections.

2.1 Linear properties of PCF

The SC-GI PCF consists of a circular core bounded by segmented cladding. Segmented cladding is characterized by the special arrangement of air holes along the radial and angular directions. $As_{38}Se_{62}$, a variant of selenide chalcogenide material is chosen as the core material. The refractive index of $As_{38}Se_{62}$ as a function of wavelength is described by Sellmeier Equation (Vyas et al. 2021):

$$n^2(\lambda) = 3.7464 + \frac{\left(3.9057 \times \lambda^2\right)}{\lambda^2 - 0.4073^2} + \frac{\left(0.9466 \times \lambda^2\right)}{\left(\lambda^2 - 40.082^2\right)} \tag{1}$$

Since the Sellmeier equation is included for defining refractive index, the involvement of both dispersion factors like material and waveguide takes place and total dispersion is calculated by the formula given as follows:

$$D(\lambda) = -\frac{\lambda}{c}\frac{d^2 n_{eff}}{d\lambda^2} \tag{2}$$

After the calculation of dispersion, Eq. (3) will provide the idea of a set of parameters related to dispersion which are evolved and termed as Taylor series expansion. In this given series, the first term of RHS defines the effective R.I. of the mode under propagation. Group wave velocity and group velocity dispersion are defined in the second and third terms of equation respectively. Consideration is taken up to the 9th order for the dispersion parameters.

$$\beta(\omega) = \beta(\omega_o) + \beta_1(\omega_o)(\omega - \omega_o) + \frac{1}{2!}\beta_2(\omega_o)(\omega - \omega_o)^2 + \frac{1}{3!}\beta_2(\omega_o)(\omega - \omega_o)^2 \ldots \tag{3}$$

2.2 Nonlinear properties of PCF

After including all the linear factors, nonlinear factors are also included for the formation of broadband SC pattern which comprises of SPM, FWM, SRS, CPM and many more. For quantifying the nonlinear effects, the nonlinear coefficient is extracted by solving the given equation:

$$\gamma = \frac{2\pi n_2}{\lambda A_{\text{eff}}} \tag{4}$$

The above equation is governed by two important specifications. One of the specifications is a nonlinear refractive index n_2. The chalcogenide glass $As_{38}Se_{62}$ has the value of n_2 as 1.13×10^{-17} m^2W^{-1} (Vyas et al. 2021). Another specification is the effective-mode-area (EMA) which is formulated below. For output of highly broadband SCG, a high magnitude of n_2 and a small EMA are favourable conditions.

$$A_{\text{eff}} = \frac{\left(\iint_{-\infty}^{\infty} |E|^2 dxdy\right)^2}{\left(\iint_{-\infty}^{\infty} |E|^4 dxdy\right)} \tag{5}$$

Incorporation of the Raman response function ($R(t\prime)$) is also done in the SC generating algorithm which includes the reliance on electronic, vibrational and nuclear contributions during the mode propagation through the PCF. Incorporation of parameters like fractional contributions, Raman period and lifetime as f_r, τ_1, τ_2 whose magnitude for material like $As_{38}Se_{62}$ are 0.1, 23.14 fs, 157 fs (Ghosh et al. 2019) are described in the following equations:

$$R(t') = (1 - f_r)\delta(t) + f_r h_r(t\prime) \tag{6}$$

$$h_r = \frac{\tau_i^2 + \tau_2^2}{\tau_1 \tau_2} \exp\left(-\frac{t\prime}{\tau_2}\right) \sin\left(\frac{t\prime}{\tau_1}\right) \tag{7}$$

2.3 Supercontinuum generation

After the interaction of all the linear and nonlinear effects which have taken participation during mode propagation, an ultra-highly spanned SC is formed. Numerically, a generalized nonlinear Schrodinger equation abbreviated as GNLSE is going to be solved with the reference of split-step Fourier Method and same is described below:

$$\frac{dA}{dz} + \frac{\alpha}{2}A - \left(\sum_{n\geq 2} \beta_n \frac{i^{n+1}}{n!} \frac{\partial^n A}{\partial t^n}\right) = i\gamma \left(1 + \frac{i}{\omega_o}\frac{\partial}{\partial t}\right) A(z,\ t) \int_{-\infty}^{\infty} R(t') \left|A(z, t - t')\right|^2 dt' \tag{8}$$

The above equation is the blend of linear and nonlinear properties organized in LHS and RHS respectively where $A(z,\ t)$ is the envelope of the optical field and α denotes the attenuation loss of $As_{38}Se_{62}$ composed PCF and having the value of 0.9 dB/m at ZDW (Le et al. 2011). All the other terms like dispersion components (β), and Raman response function $R(t')$ are already discussed in the above sections. The alterations in spectral phase across

supercontinuum are measured by relative coherence among successive spectral evolutions of SC. The first-order relative coherence is measured by the given equation:

$$\left|g_{12}^{(1)}(\lambda, t_1 - t_2)\right| = \left| \frac{\left\langle E_1^*(\lambda, t_1) E_2(\lambda, t_1) \right\rangle}{\sqrt{\left\langle \left|E_1(\lambda, t_1)\right|^2 \right\rangle \left\langle \left|E_2(\lambda, t_2)\right|^2 \right\rangle}} \right| \tag{9}$$

The numerical magnitude for coherence should be nearest to unity which verifies the degree of coherency of generated SC spectra at optimized parameters. For good agreement with SC, average coherency must be approximate to 1. Coherence must be calculated at different fiber lengths for examining the coherent nature of generated SC at maximum fiber length.

3 Designed SC-GI PCF

We have designed a Segmented Clad-Graded Index (SC-GI) photonic crystal fiber shown in Fig. 1a. The SC-GI PCF consists of a circular core bounded by segmented cladding. The radius of the solid circular core is denoted as r_c. Segmented cladding represents the low- and high-index region in angular direction (Rastogi 2001; Kumar 2012).

Here, we have realized low-index region in angular direction by incorporating three bunches of air holes whose radii are gradually increasing in radial direction. The Radii of these air holes are represented by r_{1n} (n represents the no. of the ring as clearly shown in the inset of Fig. 1a. Δr_1 represents the constant increase in radius. Λ is defined as the

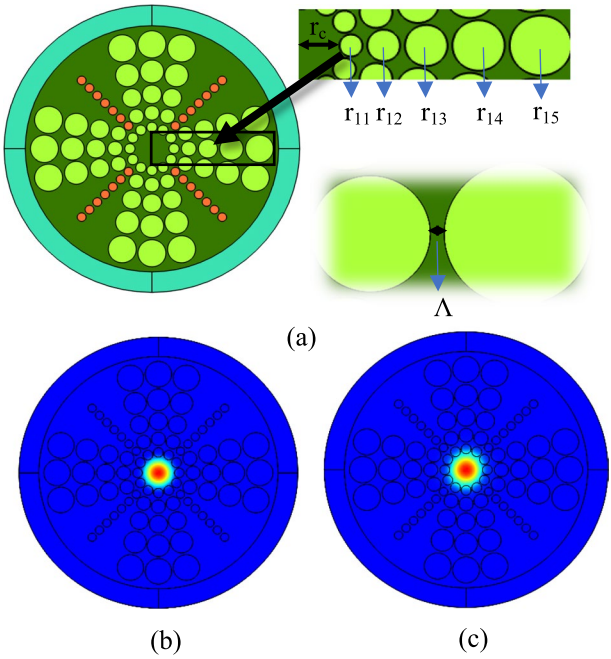


Fig. 1 **a** Proposed SC-GI PCF Design **b** Mode field at 2 μm **c** Mode field at 4 μm

constant separation between the two consecutive graded air holes as shown in the inset of Fig. 1a. Such four segmented claddings are introduced symmetrically at angles of 0, π/2, π, 3π/2 and shown in Fig. 1a. Apart from four sets of segmented cladding, another four sets of air holes of radius r_2 are placed at angles of π/4, 3π/4, 5π/4 and 7π/4. These air holes of the same radii are introduced to confine the mode properly within the core. $As_{38}Se_{62}$ chalcogenide material is used as high-index material for core and cladding region.

Apart from the PCF structure, a PML with cylindrical geometry is applied on the cross-section of fiber to eliminate the reflection back effect. We have designed our PCF to extract a nearly flat dispersion curve (with a small dispersion value) in the normal and anomalous regime. Figure 1b and c shows the fundamental modal field at 2 μm and 4 μm. In this proposed work, we have chosen anomalous region of interest which maximizes the chances of participation of soliton fission on SC.

3.1 Simulation results and discussions

This section is broken down into two sub-parts which deals with the optimization of the designed parameters (r_{11}, Δr_1 and Λ) to achieve favorable flat dispersion and hence desired broad SCG spectrum.

3.2 Evaluation of dispersion characteristics

The dispersion characteristics give the overview of the positioning of normal and anomalous regime by evaluating dispersion at broad spectral region and help in extracting the zero-dispersion wavelength (ZDW) which directly gives the information about the suitable pump wavelength that one should choose. For optimizing our parameters, first, we have seen the effect of all the designed parameters on the dispersion characteristics of the designed PCF. At the initial stage of the study, we kept our parameters at the constant value $r_{11} = 350$ nm and $\Delta r_1 = 100$ nm and varied the dimension of pitch (Λ) from 100 to 200 nm in the step of 50 nm and assessed the dispersion curves respectively. The assessment is illustrated in Fig. 2a. From the observation, we have obtained two values of zero dispersion wavelength in 2000–7000 nm spectral wavelength. The first ZDW is found around 2300 nm which keeps on increasing on the increment of Λ and another ZDW which falls around 6300 nm is on the verge of decrement in the same set of observations. Next, we have seen the effect of r_{11} while keeping $\Delta r_1 = 100$ nm and $\Lambda = 150$ nm and shown in Fig. 2b.

On the increment of r_{11}, both the ZDWs shift towards high wavelengths (red-shifted). After that, we have taken $r_{11} = 350$ nm and $\Lambda = 150$ nm as constant parameters and changed the grading (Δr_1) from 100 to 200 nm. We have noticed that there is no significant change in the first ZDW but the increment is observed in the second ZDW (Fig. 2c). At the next step of observation we have influenced the radius of circular core denoted as r_c and changed the magnitude from 900 to 1100 nm at a step of 100 nm. The observation displayed in Fig. 2d shows that first ZDW got red shifted and increased by 100 nm at increment of the r_c. The respective second ZDW got blue shifted by the same influence. After that, we introduced an increment in the magnitude of r_2 by 50 nm from 250 to 350 nm. The observation shown in Fig. 2e concludes that there is no change in first ZDW, whereas second ZDW experienced a negligible change. In the final step of study, we have varied the angle of rotation of three bunches of cladding holes by 45° and no variation has been observed in dispersion characteristics displayed in Fig. 2f because of the symmetricity of the structure. With the above analysis, we have fixed our optimized design parameters as

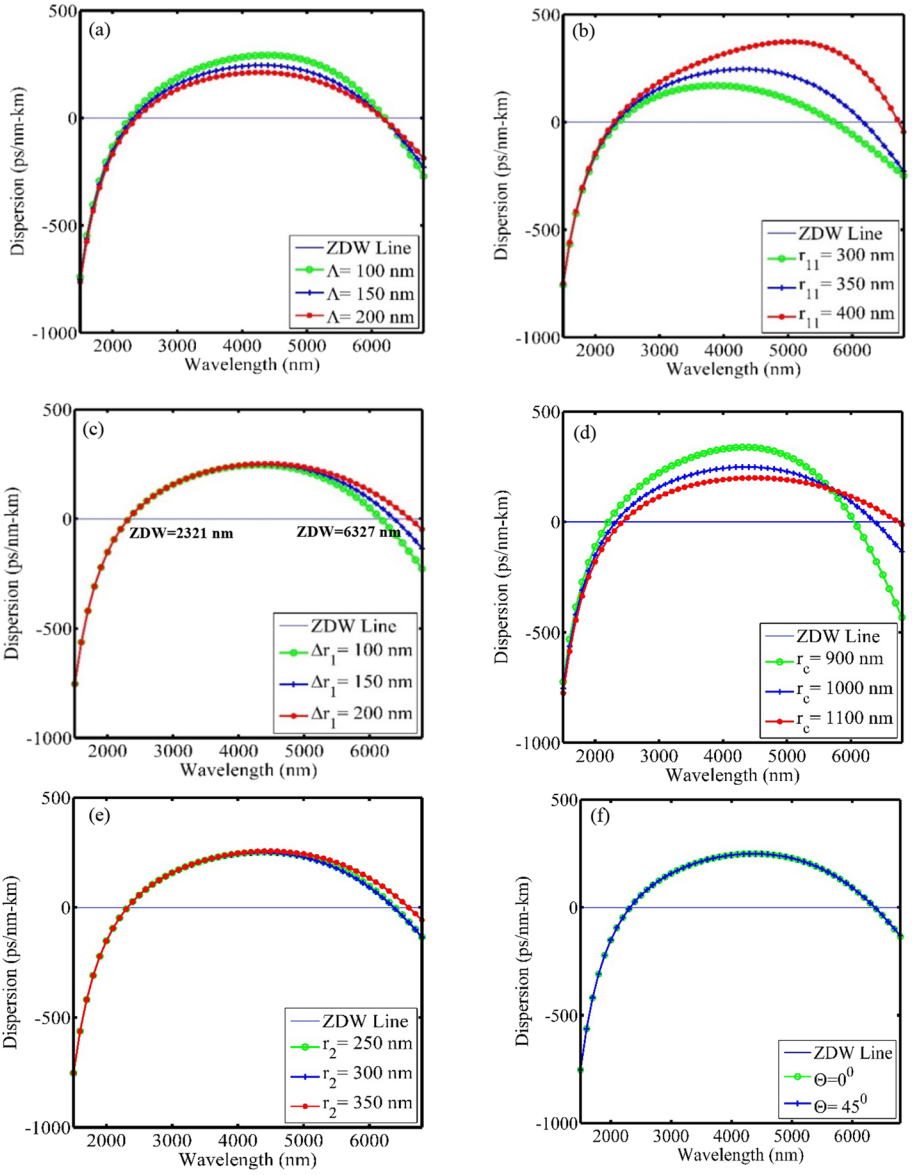


Fig. 2 Dispersion Characteristics with variation of **a** Pitch (Λ) **b** Radius (r_{11}) **c** Grading (Δr_1) **d** Radius of circular core (r_c) **e** Radius (r_2) **f** Angle of rotation of segmented cladding (Θ)

shown in Table 1. The reason behind the optimization is the selection of ZDW. We have extracted values of ZDW as 2321 nm and 6367 nm at our optimized parameters which helps us to choose our pump wavelength as 3.4 μm laser source (anomalous region) which is readily available OPCPA Laser (Petrov 2001; Mayer et al. 2013).

Figure 3 shows the spectral variation of EMA and the nonlinear coefficient. EMA keeps on increasing with wavelength due to the weak confinement of light at higher wavelengths.

Table 1 Optimized design parameters

	Pitch (Λ)	Radius (r_{11})	Change in radius (Δr_1)	Radius of circular core (r_c)	Radius (r_2)	Segmented Clad's angle of rotation (Θ)
PCF parameters	150 nm	350 nm	150 nm	1000 nm	300 nm	0°

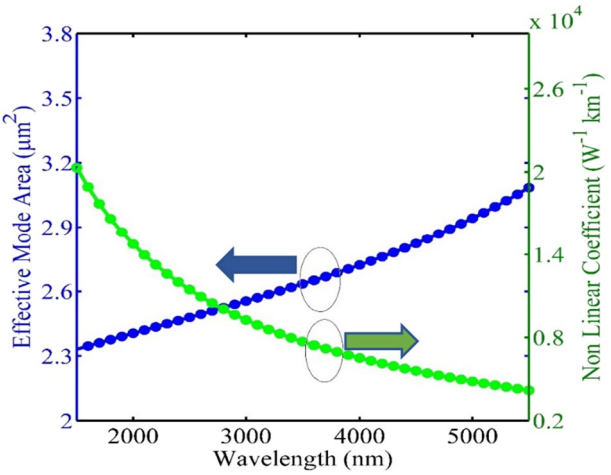


Fig. 3 Variation of effective mode area and nonlinear coefficient with respect to wavelength at optimized parameters

In response to this, the nonlinear coefficient that inversely depends on A_{eff} keeps on decreasing. At pump wavelength $\lambda = 3.4$ μm, the proposed fiber (at optimized parameters) shows a small magnitude of A_{eff} (2.46 μm^2) and a correspondingly large nonlinear coefficient attaining (12,005 W^{-1} km^{-1}).

3.3 Evolution of supercontinuum at anomalous region

For the formation of broadband SC, a secant hyperbolic laser pulse is fed to the PCF containing all the optimizing parameters. Numerically, we have solved GNLSE already discussed in Sec. II. B. with the help of MATLAB software at different peak powers (kW), pulse widths (femtoseconds) and different fiber lengths (mm). The appropriate action of the nonlinearities participating in spectral broadening is clarified by calculated values of the distinct time scales tabulated in Table 2

During the pumping of femtosecond pulse at 3.4 μm which falls at anomalous region, an ultra-broadband SC spectrum is generated at relatively low peak powers 1.8 kW which are displayed in Fig. 4.

Table 2 Length scales and other nonlinear parameters

Length scale/parameter	Formula	Calculated value
Peak power	P_0	1.8 kW
Pulse width	T_{FWHM}	50 fs
Pulse duration	$T_0 = \frac{T_{FWHM}}{1.763}$	28.3 fs
Group velocity dispersion	β_2	1.37 ps^2/m
Nonlinear length	$L_{NL} = \frac{1}{\gamma P_0}$	0.0463 mm
Dispersion length	$L_D = \frac{T_0^2}{\lvert\beta_2\rvert}$	0.585 mm
Soliton order	$N = \sqrt{\frac{L_D}{L_{NL}}}$	≈ 4
Soliton fission length	$L_{fs} = \frac{L_D}{N}$	0.1463 mm

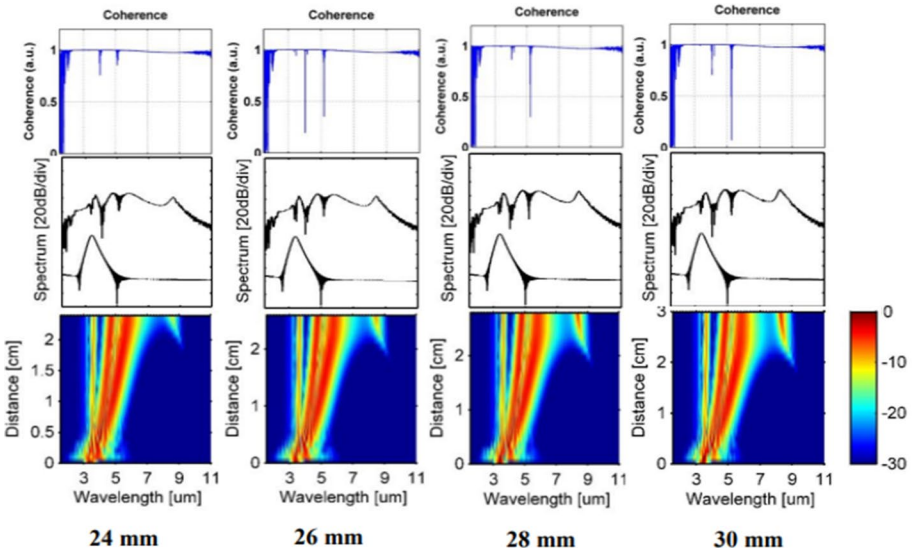


Fig. 4 Numerically simulated SC spectra at the pumping of 3.4 µm at different fiber lengths

Figure 4 shows the change in SC broadening with fiber length. SCG bandwidth increases as fiber length is increased from 24 to 28 mm. Due to the small effective mode area and higher nonlinear coefficient, nonlinear effects like SPM and XPM participate very early in respective SC broadening at millimeter scale of length. In quantitative speaking, nonlinear effects started to participate after pulse propagation around 0.05 mm (because $L_{NL}=0.0463$ mm) and dispersive effects begin to participate after propagating 0.6 mm (because $L_D=0.585$ mm) along fiber length. Since we have worked on anomalous region, soliton fission is the main attraction for SC broadening alongwith SPM and XPM. At anomalous dispersion, there exhibits some instability in the nonlinear medium which modulates the steady state in such a way that gives birth to train of ultrashort pulses known as solitons which remain unaffected in propagation of longer lengths. Due to the undistorted nature of solitons, they got entangled at certain length and generate new frequency components in lower frequency range which basically defined as soliton fission. The soliton order is (~4) and the respective soliton fission length (L_{fs}) is 0.15 mm. Raman Scattering and higher order dispersion are the main two factors that can perturb periodic evolution ideally and respective induction of pulse breakdown with the aid of soliton fission. But further increment in the fiber length does not influence the broadening of the SC spectra. Thus, we have taken 28 mm as an optimum fiber length to achieve the favorable SCG bandwidth spanning from 3000 to 9500 nm which fully wraps the transparency region at very low peak power of 1.8 kW.

Figure 5 shows the SC spectrum at variable peak powers (1.6 kW to 1.8 kW) with a constant fiber length of 28 mm and pulse width of 50 fs. The quantitative understanding from nonlinear length given in Table 2 is gathered as increment in peak power results in introduction of nonlinear effects faster at a small length of the fiber. Therefore, a significant increase in the SCG bandwidth is observed with the increase in peak power till 1.8 kW. It has also been observed that further increase in peak power (> 1.8 kW) is not showing any significant improvement in the SCG bandwidth and the coherence of

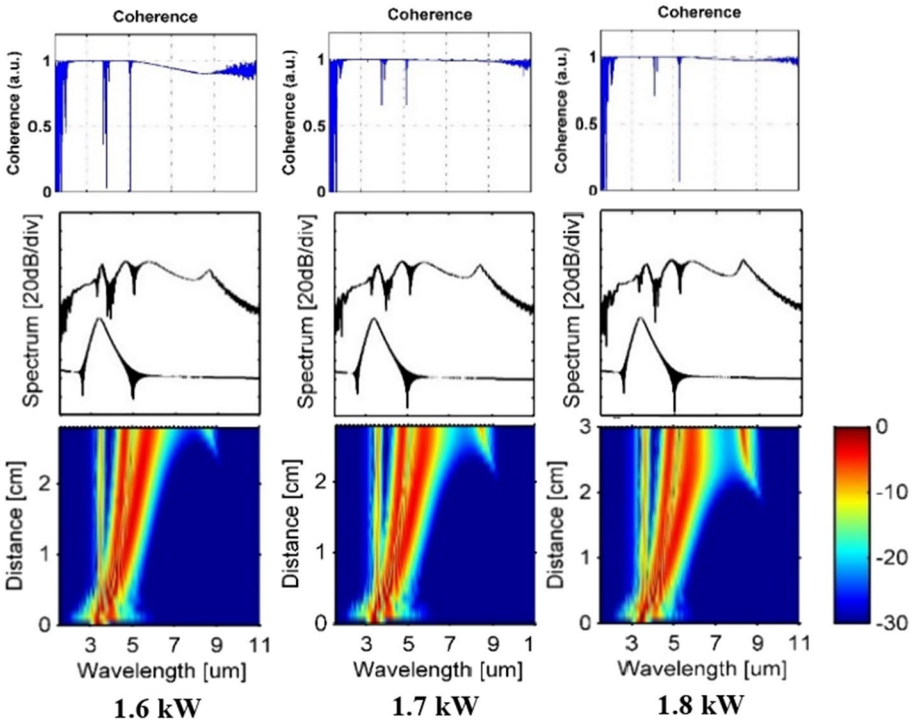


Fig. 5 Numerically simulated SC spectra at the pumping of 3.4 μm at different peak powers

the spectrum is also get disturbed. Thus, we have kept 1.8 kW peak power as optimum peak power of input pulse for analysis. The effect of pulse width (50 fs to 90 fs) on the SCG spectrum is shown in Fig. 6. Since, increase in pulse width (T_{FWHM}) and respective increase in pulse duration (T_0) causes the delay in nonlinear effects by which an intense broadband spectrum is not visible with the increment of the pulse width. Either the band

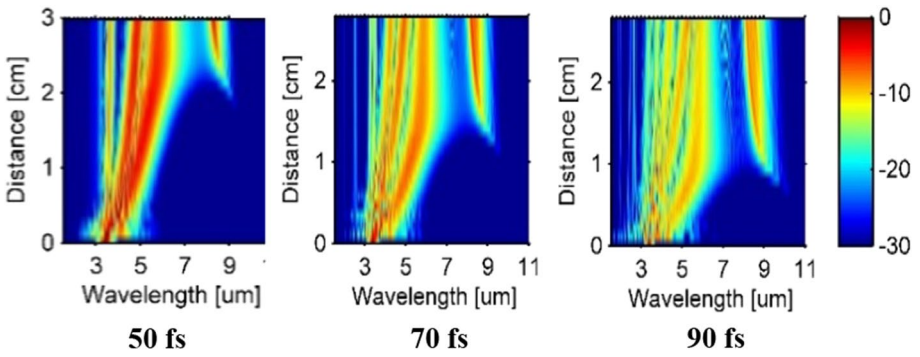


Fig. 6 Spectral Evolution at different pulse width

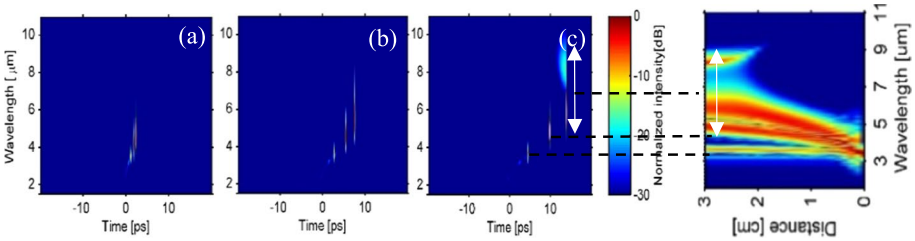


Fig. 7 Time-Wavelength FROG spectrogram at **a** 8 mm **b** 18 mm **c** 28 mm

is spanning in red-shifted side but a poor intensity (approximate half amplitude) or distorted coherence pattern is observed with pulse width.

Figure 7 displayed the time-wavelength relationship known as FROG spectrogram at the desired laser input fed to different lengths of fiber. The FROG spectrogram helps to identify the dispersive components along with soliton fission components distinctly and also tells about their mutual correlation. Time-delayed spectral components are observable at respective spectrograms.

The proposed PCF can be fabricated by the methods of extrusion and drilling techniques. The extrusion method offers an advantage over drilling techniques because the extrusion approach can tolerate high dopant concentrations, which provide a route, it offers greater flexibility in the construction of PCFs made from chalcogenide glasses to more readily change the mechanical and optical characteristics in contrast to silica glass (Chauhan 2020; Zhang 2015; Ruan 2010).

From all the outputs, we can conclude that a broad spectral evolution of SC of 3000–9500 nm is generated at 1.8 kW of peak power, when pumped by 3.4 μm femtosecond laser of 50 kHz repetition rate and pulse width of 50 fs is fed to 28 mm proposed PC fiber. In comparison to previous works as described in Table 3, our PCF can deliver broad SC spectra at relatively low peak power and with small fiber length.

4 Conclusion

We have devised for the first time SCG analysis in Segmented Cladding Design in the author's knowledge. The presented The SC-GI PCF consists of a circular core bounded by segmented cladding. The low-index region in angular direction of the Segmented Cladding is achieved by a specially designed air holes whose diameters increases gradually achieved and high-index regions is realized by $As_{38}Se_{62}$ chalcogenide material. The presented novel fiber design offers two ZDWs at 2321 nm and 6367 nm at optimized design parameters. Correspondingly we have chosen 3.4 μm as our pump wavelength which falls in anomalous regime. At revamped parameters shown in Table 2, our PCF results in achieving a very small effective mode area having magnitude 2.46 μm^2 and high nonlinear coefficient of 12,005 W^{-1} km^{-1}. The Supercontinuum spanning from 2 to 9.5 μm at 3.4 μm pumping at level of − 30 dB is reported which completely covers the transparency limit of proposed material and also covers the Mid-IR region in some extant. Applications like optical metrology, optical coherence tomography (OCT), food quality control & telecommunications has the need of broadband SC spectra which we have simulated in this work.

Table 3 Comparison of our results with previously done works

Sr. No	Material used	SCG Spectra	Broadening	PeakPower & Fiber length	References
1	As_2Se_3	2.53–16.6 µm	14 µm	4000 W & 47 mm	Feng et al. (2020)
2	$Ge_{12}As_{24}Se_{64}$-$Ge_{10}As_{24}S_{66}$	1.8–10 µm	8.2 µm	3000 W & 3.61 cm	Yu et al. (2013)
3	$As_{39}Se_{61}$	1.95–6.58 µm	4.63 µm	880 W & 1 cm	Medjouri and Abed (2019)
4	$Ge_{10}As_{22}Se_{68}$-As_2S_3	2–8 µm	6 µm	5000 W & 10 mm	Karim (2020)
5	As_2S_5	0.7–5.2 µm	4.5 µm	25,000 W & 5 mm	Diouf (2019a, b)
6	$As_{38}Se_{62}$	2–10 µm	8 µm	1330 W & 200 mm	Vyas et al. (2021)
7	$As_{38}Se_{62}$	3–9.5 µm	6.5 µm	1800 W & 28 mm	This work

Author contributions A Khamaru has done the simulation work, analyses the results and written the original manuscript and A Kumar has given ideas and review the manuscript.

Funding No funding is provided for this research.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Alfano, R.R., Shapiro, S.L.: Emission in the region 4000 to 7000 Å via four-photon coupling in glass. Phys. Rev. Lett. **24**, 584 (1970)

Chauhan, P., Kumar, A., Kalra, Y.: Numerical exploration of coherent supercontinuum generation in multicomponent $GeSe_2$-As_2Se_3-PbSe chalcogenide based photonic crystal fiber. Opt. Fiber Technol. **54**, 102100 (2020). https://doi.org/10.1016/j.yofte.2019.102100

Chu, V., et al.: Multi-octave supercontinuum generation in As_2Se_3 chalcogenide photonic crystal fiber. Phot. Nanostruct. (2022). https://doi.org/10.1016/j.photonics.2021.100986

Diouf, M., Wague, A., Zghal, M.: Numerical investigation of an ultra-broadband coherent mid-infrared supercontinuum in a chalcogenide $AsSe_2$-As_2S_5 multimaterial photonic crystal fiber. J. Opt. Soc. Am. B **36**, A8–A14 (2019a). https://doi.org/10.1364/JOSAB.36.0000A8

Diouf, M., Mandeng, L.M., Tchawoua, C., Zghal, M.: Numerical investigation of supercontinuum generation through $AsSe_2$/As_2S_5 chalcogenide photonic crystal fibres and rib structures. J. Lightw. Technol. **37**, 5692–5698 (2019b). https://doi.org/10.1109/JLT.2019.2934034

Dussardier, B., Rastogi, V., Kumar, A., Monnom, G.: Large-mode-area leaky optical fiber fabricated by MCVD. Appl. Opt. **50**, 3118–3122 (2011). https://doi.org/10.1364/AO.50.003118

Feng, X., et al.: Ultra-large mode area chalcogenide photonic crystal fiber for high-power mid-infrared applications, In: 2020 Conference on lasers and electro-optics (CLEO), pp. 1–2 (2020)

Ghosh, A.N., et al.: Chalcogenide-glass polarization-maintaining photonic crystal fiber for mid-infrared supercontinuum generation. J. Phys.: Photon. **1**, 44003 (2019). https://doi.org/10.1088/2515-7647/ab3b1e

Hartl, I., et al.: Ultrahigh-resolution optical coherence tomography using continuum generation in an air–silica microstructure optical fiber. Opt. Lett. **26**, 608–610 (2001). https://doi.org/10.1364/OL.26.000608

Jiao, K., et al.: 12–15.2 μm supercontinuum generation in a low-loss chalcohalide fiber pumped at a deep anomalous-dispersion region. Opt. Lett. **44**, 5545 (2019). https://doi.org/10.1364/ol.44.005545

Karim, M.R., Kayed, N., Dey, G.K., Rahman, B.M.A.: Design and analysis of suspended core channel waveguide made using As_2Se_3 glass system for mid-infrared supercontinuum generation. J. Opt. **23**, 155 (2020). https://doi.org/10.1088/2040-8986/abcfd3

Khamaru, A., Sharma, K., Tomer, D.S., Kumar, A.: Mid-IR supercontinuum generation in dispersion flattened As38Se62 chalcogenide photonic crystal fiber. J. Phys. Conf. Ser. **2426**, 12002 (2023). https://doi.org/10.1088/1742-6596/2426/1/012002

Kim, M., Lee, C., Kim, S.: Silicon-embedded photonic crystal fiber for high birefringence and nonlinearity. Optik **212**, 164657 (2020). https://doi.org/10.1016/j.ijleo.2020.164657

Kumar, A., Rastogi, V., Rahman, B.M.A.: Birefringence analysis of segmented cladding fiber. Appl. Opt. **51**(15), 3104–3108 (2012). https://doi.org/10.1364/AO.51.003104

Le, S.D., et al.: Efficient four-wave mixing in an ultra-highly nonlinear suspended-core chalcogenide $As_{38}Se_{62}$ fiber. Opt. Expr. **19**, B653–B660 (2011). https://doi.org/10.1364/OE.19.00B653

Liao, M., Chaudhari, C., Qin, G., Yan, X., Suzuki, T., Ohishi, Y.: Tellurite microstructure fibers with small hexagonal core for supercontinuum generation. Opt. Expr. **17**, 12174–12182 (2009). https://doi.org/10.1364/OE.17.012174

Matsui, T., Zhou, J., Nakajima, K., Sankawa, I.: Dispersion-flattened photonic crystal fiber with large effective area and low confinement loss. J. Lightw. Technol. **23**, 4178–4183 (2005). https://doi.org/10.1109/JLT.2005.858250

Mayer, B.M., et al.: Sub-four cycle laser pulses directly from a high-repetition-rate optical parametric chirped-pulse amplifier at 3.4 μm. Opt. Lett. **38**, 4265–4268 (2013). https://doi.org/10.1364/OL.38.004265

Medjouri, A., Abed, D.: Design and modelling of all-normal dispersion $As_{39}Se_{61}$ chalcogenide photonic crystal fiber for flat-top coherent mid-infrared supercontinuum generation. Opt. Fiber Technol. **50**, 154–164 (2019). https://doi.org/10.1016/j.yofte.2019.03.021

Medjouri, A., Abed, D.: Mid-infrared broadband ultraflat-top supercontinuum generation in dispersion engineered Ge-Sb-Se chalcogenide photonic crystal fiber. Opt Mater (Amst) (2019). https://doi.org/10.1016/j.optmat.2019.109391

Medjouri, A., Abed, D.: Modelling of all-chalcogenide all-normal dispersion photonic crystal fiber for ultra-flat mid-infrared supercontinuum generation. Opt. Quant. Elec. **53**, 1–22 (2021). https://doi.org/10.1007/s11082-021-03027-2

Møller, U., et al.: Multi-milliwatt mid-infrared supercontinuum generation in a suspended core chalcogenide fiber. Opt. Expr. **23**(3), 3282–3291 (2015). https://doi.org/10.1364/OE.23.003282

Petrov, V., Rotermund, F., Noack, F.: Generation of high-power femtosecond light pulses at 1 kHz in the mid-infrared spectral range between 3 and 12 μm by second-order nonlinear processes in optical crystals. J. Optics A: Pure Appl. Optics (2001). https://doi.org/10.1088/1464-4258/3/3/201

Rastogi, V., Chiang, K.S.: Propagation characteristics of a segmented cladding fiber. Opt. Lett. **26**, 491–493 (2001). https://doi.org/10.1364/OL.26.000491

Ruan, Y., Ebendorff-Heidepriem, H., Afshar, S., Monro, T.M.: Light confinement within nanoholes in nanostructured optical fibers. Opt. Expr. **18**, 26018–26026 (2010). https://doi.org/10.1364/OE.18.026018

Saini, T.S., Kumar, A., Sinha, R.K.: Broadband mid-IR supercontinuum generation in As_2Se_3 based chalcogenide photonic crystal fiber: a new design and analysis. Opt. Commun. **347**, 13–19 (2015). https://doi.org/10.1016/j.optcom.2015.02.049

Saini, T.S., Kumar, A., Sinha, R.K.: Design and modelling of dispersion-engineered rib waveguide for ultra broadband mid-infrared supercontinuum generation. J. Mod. Opt. **64**, 143–149 (2017). https://doi.org/10.1080/09500340.2016.1216190

Salimullah, S., Faisal, M.: Ultra-wideband and coherent supercontinuum generation (near and mid-infrared) in dispersion flattened $ZnGeP_2$ photonic crystal fiber. Alex. Eng. J. **70**, 289–300 (2023). https://doi.org/10.1016/j.aej.2023.03.002

She, Y., Zhang, W., Tu, S., Liang, G.: Large mode area single mode photonic crystal fiber with ultra-low bending loss. Optik **229**, 165556 (2021). https://doi.org/10.1016/j.ijleo.2020.165556

Siwach, P., Kumar, A., Saini, T.S.: Broadband supercontinuum generation spanning 1.5–13 μm in $Ge_{11.5}As_{24}Se_{64.5}$ based chalcogenide glass step index optical fiber. Optik **156**, 564–570 (2018). https://doi.org/10.1016/j.ijleo.2017.11.199

Vyas, S., Tanabe, T., Tiwari, M., Singh, G.: Chalcogenide photonic crystal fiber for ultraflat mid-infrared supercontinuum generation. Chinese Optics Lett. **14**, 123201–123205 (2016). https://doi.org/10.3788/col201614.123201

Vyas, S., et al.: Mid-infrared supercontinuum generation using low peak pump power in $As_{38}Se_{62}$ based chalcogenide photonic crystal fiber. IOP Conf Ser Mater Sci Eng **1119**(1), 012022 (2021). https://doi.org/10.1088/1757-899X/1119/1/012022

Wang, X., et al.: Mid-infrared supercontinuum generation in low-loss single-mode Te-rich chalcogenide fiber. Opt. Mater. Expr. **9**(8), 3487 (2019). https://doi.org/10.1364/ome.9.003487

Wang, X., Lei, H., Xie, K., Li, J.: All-silica fiber ultra-flat 2–3 μm supercontinuum source based on highly nonlinear silica fiber. IEEE Phot. Technol. Lett. **35**, 381–384 (2023). https://doi.org/10.1109/LPT.2023.3245602

Xu F et al.: Mid-infrared self-similar pulse compression in a tapered tellurite photonic crystal fiber and its application in supercontinuum generation. J. Lightwave Technol., **36** 3514–3521 (2018) https://opg.optica.org/jlt/abstract.cfm?URI=jlt-36-16-3514

Ye, J., Schnatz, H., Hollberg, L.W.: Optical frequency combs: from frequency metrology to optical phase control. IEEE J. Sel. Top. Quantum Electron. **9**, 1041–1058 (2003). https://doi.org/10.1109/JSTQE.2003.819109

Yu, Y., et al.: Mid-infrared supercontinuum generation in chalcogenides. Opt. Mater. Expr. **3**, 1075–1086 (2013). https://doi.org/10.1364/OME.3.001075

Zhang, P., et al.: Fabrication of chalcogenide glass photonic crystal fibers with mechanical drilling. Opt. Fiber Technol. **26**, 176–179 (2015). https://doi.org/10.1016/j.yofte.2015.09.002

Zheng, J., et al.: Optical nonlinearity and supercontinuum generation of tellurite glass at 1.064 μm. Opt Laser Technol **138**, 106832 (2021). https://doi.org/10.1016/j.optlastec.2020.106832

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT) | 979-8-3503-7131-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/InCACCT61598.2024.10550990

BEACOMP: A Novel Textual Adversarial Attack Architecture for Unveiling the Fragility of Neural Text Classifiers

Rishabh Sharma
Department of Electrical Engineering
Delhi Technological University (DTU)
Delhi, India
rsharma5752@gmail.com

Aatish Ghosh
Department of Electrical Engineering
Delhi Technological University (DTU)
Delhi, India
aatish.ghosh@gmail.com

Ashish Bajaj
Department of Information Technology
Delhi Technological University (DTU)
Delhi, India
bajaj.ashish25@gmail.com

Dinesh Kumar Vishwakarma
Department of Information Technology
Delhi Technological University (DTU)
Delhi, India
dvishwakarma@gmail.com

***Abstract*—Natural Language Processing models are now being used more than ever before and they have been quite remarkable in applications like sentiment classification, text generation, generative AI, etc. The deep learning models behind these applications of Natural Language Processing are the reason why these models can understand the complex relation between the variables and apply those learning for the applications of NLP. But, similar to all the other applications of deep learning, NLP models are also equally susceptible to external attacks which can hamper the accuracy and the credibility of the model, which in the end poses a big question on the reliability of these models. Textual adversarial attacks can deviate the model from giving the correct answer to giving an incorrect output, just because of a small perturbation made in the input. These perturbations, often involving synonyms of the original word, are still able to successfully fool these models. This paper, in a similar context, proposes BEACOMP, a novel textual adversarial attack that uses composite transformation functions from the TextAttack library, in a black box setting, to successfully fool the model. The accuracy and the success rate of the attack are validated through three different models Distilled BERT, BERT uncased, and RoBERTa, on 4 different datasets including IMDB, SST2, Rotten Tomatoes, and AG news. The use of the attack proposed showed an increase in the attack success rate, which in the end improves the model's preparedness for adversarial attacks. On average, for the models tested in the research, their accuracy dropped from 97.16% to 1.91%. This demonstrates the effectiveness of the TextAttack library in generating attack patterns for text classification patterns while adhering to certain constraints.**

***Index Terms*—Natural Language Processing, Adversarial Machine Learning, Robustness of AI Models, Deep Learning, TextAttack**

I. INTRODUCTION

Applications of NLP models are endless in today's era, ranging from sentiment analysis, chatbots, machine translation, etc. to the latest developments in the AI world like generative AI and LLMs. Products like ChatGPT and BardAI have expressed the significant impact NLP can make in today's world of deep learning. The acceptability and awareness of NLP models being used in our daily lives have been increasing significantly day by day, in which the LLM models discussed above have a substantial part to play. It is evident that with the rapid increase in the utilization of these NLP models, the threat to these models is an issue that needs to be tackled while working on them. These models, if not calibrated for these adversarial attacks [1] carefully, can not only be an issue in the accuracy of these models but more importantly can be a massive pool of misinformation. With small perturbations, these models can be fooled very easily which poses serious concerns over the reliability of the models [2]. If compared to a real-life scenario, with the ever-increasing popularity and usage of LLM models, with changes as small as character level changes in the input, the algorithm behind the LLMs can be easily fooled and can give a completely different answer than what should be the correct answer [3]. One of the problems is also that most of the attacks which have been discussed in the context of TextAttack have been single parametric attacks, i.e, little to no research has been done on the possibility of using composite attacks combining different parameters of the four pillars of the TextAttack library(Goal function, Transformation Function, Constraint Definition, and Search Method Selection), into a single attack. This possibility has several possible advantages over single parametric attacks which have been discussed before. Extending on the possibility of increased perturbations, as there will be a better quality adversarial example generated, the Attack Success rate will be increased as more words will be changed.

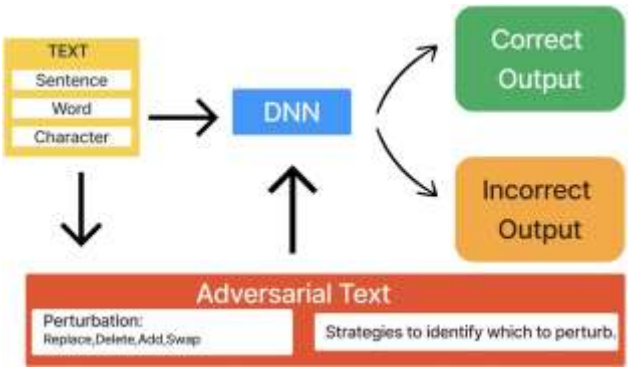


Fig. 1. Visual depiction of a general adversarial attack

979-8-3503-7131-4/24/$31.00 ©2024 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 20,2024 at 06:19:33 UTC from IEEE Xplore. Restrictions apply.

The purpose of the study is to create a **novel attack** using the TextAttack library which covers a large number of words and symbols so that a more robust attack function could be created. It also allows the attack to be used in a variety of applications pertaining to different industries [4] as the attack function is capable of generating adversarial examples which are very effective with a high attack success rate and a high average perturbed rate.

As shown in Figure 1, one of the key areas of adversarial machine learning is untargeted classification in NLP models and this study dis- cusses a novel adversarial attack 'BEACOMP' which in a nutshell uses beam search and a composite transformation function for untargeted classification of various NLP models like RoBERTa, Distilled BERT, BERT Uncased, etc. The unique features of the above adversarial attack are discussed below:

A. Composite Transformation Function

The previous works in the area have used a single trans- formation function to run the attacks whereas in this study, a combination of various transformation functions is used along with Beam Search to enhance the efficiency of the attack on the NLP models. Using a composite transformation function rather than a single transformation function increases the vari- ability of usage of the attack as the composite transformation function involves perturbations using WordSwap Embedding, Word Deletion, WordSwap Change Number, etc.

B. Attack Success Rate

Due to the uniqueness of the transformation function, an increase in success rate was seen. This can also be held accountable due to the fact that even the embedding distance was not kept less than 0.7 so as to not lose the semantic meaning of an input, therefore, the attack was able to perform better and hence the attack success rate was seen to be increased as compared to the precious studies.

C. Usability

Due to the seamless integration of the text attack library with various pre-trained models, it makes it very efficient to implement BEACOMP with various NLP models, even the models and datasets other than those discussed in this study. This makes the potential usage of BEACOMP attack for untargeted classification very seamless with various NLP models.

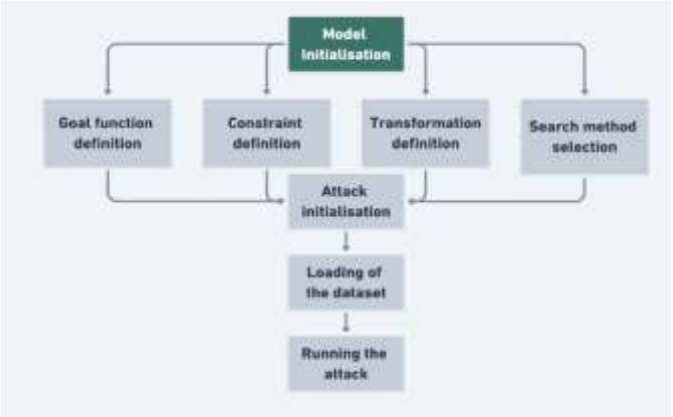


Fig. 2. Proposed Approach of the Attack

II. PROPOSED APPROACH

BEACOMP is a novel architecture designed to launch tex- tual adversarial attacks, which aim to expose the vulnerability of neural text classifiers. As shown in Figure 2, the approach for the BEACOMP attack can be described in 7 different steps ranging from Model initialisation, Goal function definition, Constraint definition, Transformation definition, Search method selection, Attack initialization, loading of the dataset, and then finally running the attack. The figure 2 clearly shows the step-by-step process
[5] of the entire attack as the steps of initialising the attack specifications:

A. Model Initialisation

The attack initialization starts with deciding on a pre- trained model from the TextAttack library for sequence classification. Along with the model, a tokenizer is also initialized for the attack according to the model selected, using the HuggingFace transformers library. Both of these entities are then wrapped in an interface named model wrapper using the 'Hugging- FaceModelWrapper'. This wrapper acts as a bridge, allowing seamless integration of th pre-trained BERT model [6] into the TextAttack framework for text-based attacks and evaluations.

B. Goal Function Definition

The goal function explains the target of our attack, based on our choice, The goal function can be divided into two parts: the first one is Untargeted Classification, in which our primary goal is perturbation the input just enough to deviate the output to anything but the original correct output. The other one is targeted classification where we are responsible for dictating the output classification label of the attack, and the output is controlled by the attack. For this study, we have focused on using Untargeted Classification for our attack.

c) Constraint Definition

The constraints are one of the most vital aspects of the attack as they basically outline the transformation limits of the input text so as to semantically make sense and be as close to the input text as possible. For this research work, a combination of 3 different constraints has been used so as to cover the maximum number of constraints as possible.

a. Repeat Modification

Repeat Modification constraints the modification of words which have already been modified. The RepeatModification Constraint tracks the words which have already been modified so as to prevent unnatural and irrelevant responses and also to save on computational time.

b. StopWordModification

StopWords in the context of NLP are defined by words which are not semantically relevant to the sentence for its understanding and hence are omitted when performing modifications in the sentence. StopWordModification constraint does exactly this, i.e., prevents the modification of stopwords in a sentence.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 20,2024 at 06:19:33 UTC from IEEE Xplore. Restrictions apply.

c. WordEmbeddingDistance

As shown in Figure 3, the WordEmbeddingDistance uses the concept of 'Cosine Similarity' [7] to determine how sim- ilar words are by comparing the cosine angle between their vectors in multi-dimensional space. As shown in Figure 4, this helps in finding the most semantically relevant words for the modification to the input text. Now, the WordEmbed- dingDistance should generally not have a low value as that will produce irrelevant words as the modification. On the other side, the WordEmbeddingDistance cannot be very high either as that will make the process of finding a word which is semantically similar yet different very difficult. Hence, for this research study, we took the WordEmbeddingDistance to be equal to or greater than 0.7.

D. Transformation Definition

The transformation functions defined in each attack are the main functions which control the transformation of the input text and the type of perturbations allowed with the input. As shown in Figure 5, These modifications may be in the form of numbers in case of numerical input or alphabetical modifications in case of textual input and the types of transformations allowed are guided by the transformation functions defined in the attack.

As shown in figure 6, For this research work, a composite function of 3 transfor- mation functions has been used with the functions being used as follows:

a. WordSwapEmbedding

As discussed in the section on WordEmbeddingDistance constraint, the semantically nearest according to the Word embedding [8] distance is then swapped with the original word. The swapping takes place according to the word embedding distance defined in the WordEmbeddingDistance constraint definition and only then the modification of the original input takes place with perturbation.

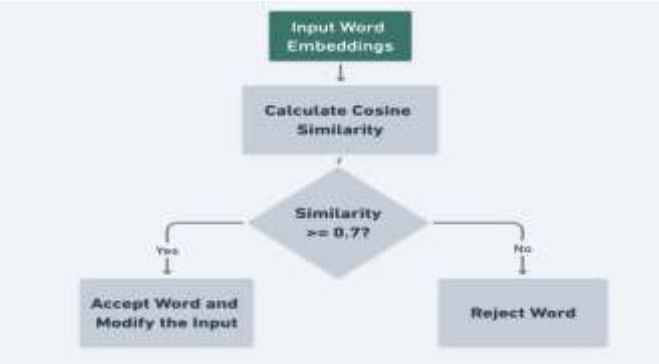


Fig. 3. Calculating world Embedding distance.

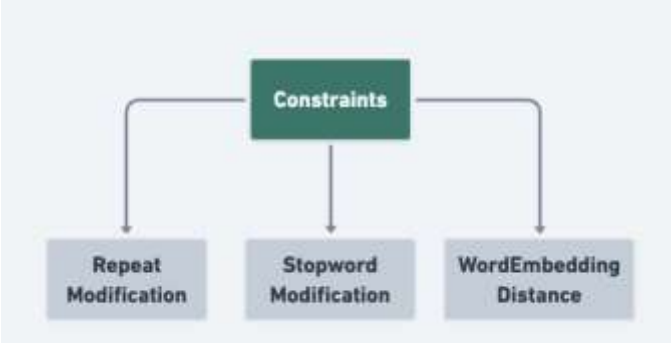


Fig. 4. Characteristics of the Constraints put on novel attack.

b. Word Deletion

Word Deletion is a transformation method that takes a sentence and transforms it by deleting a single word. This transformation function helps in increasing the randomness of the input text as there are no pre-defined criteria as to which word will get deleted, yet retaining the semantic structure and the meaning of the sentence.

c. WordSwapChangeNumber

More often than not, inputs involve numbers as part of the sentence and the WordSwapChangeNumber is used to change any random number in the sentence to make a valid perturbation input. One of the key examples in the classification space for this transformation could be movie ratings, for example, if "I give 1.5 out of 5 stars to the movie" is changed to "I give 4.5 out of 5 stars to the movie", this makes a valid perturbed sentence which will dis-classify the output.

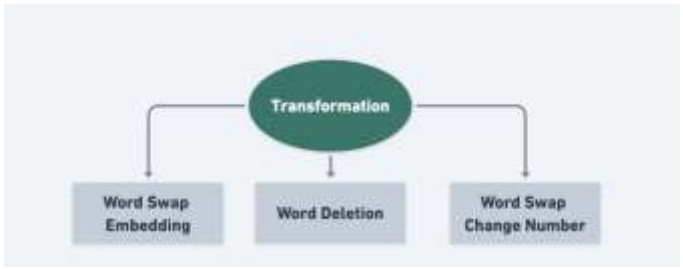


Fig. 5. Characteristics of Transformations put on novel attack.

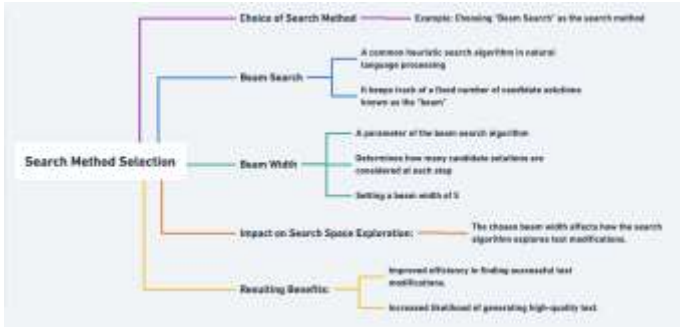


Fig. 6. Characteristics of Transformations put on novel attack.

E. Search Method Selection

For the context of this research, Beam-Search was used as the search method for searching possible text modifica- tions during the attack. BeamSearch [9] was used instead of GreedySearch as BeamSearch, even if it takes a higher computational power, produces more accurate results.

The algorithm uses a parameter called "beamwidth" that determines the number of adjustments required at each step of the search. For example, choosing a beam width of 5 means that the algorithm searches for and evaluates up to five different text editors simultaneously. This approach increases the risk of effectively detecting text changes by effectively searching the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 20,2024 at 06:19:33 UTC from IEEE Xplore. Restrictions apply.

search space. In general, the choice of the search method and its settings affect the effectiveness and efficiency of creating enemy attacks in the TextAttack library [10]. For our research, the beam width was kept to be 5.

F. Attack Initialisation

The novel attack is then initialized with all the parameters i.e., Goal Function, Constraints, Transformations, and Search method, written in the definition part of the attack. An object of 'attack' is then created containing all the parameters. The attack encapsulates the entire attack configuration which will be applied to the datasets, the process of which, is discussed in the next section.

G. Loading of the Dataset

The next process comprises deciding the dataset on which the attack will be tested. For this study, datasets which have been used over here have variety in terms of having binary output, multiple classification output, etc. Some of the exam- ples include sentiment classification as positive or negative, or Deciding the category of the news according to the text as sci-tech, business, or world news, etc.

Nonetheless, in both the types of the above datasets, untar-geted classification was used and the goal was to use famous models like Distilled BERT, BERT uncased, and RoBERTa on these datasets to testify to the accuracy and efficiency of the attack [11]. The datasets which were used for this study were described in Table 1.

TABLE I. DATASET DESCRIPTION AND OUTPUT TYPE

Dataset Name	Description	Output Type
AG News	AG is a collection of more than 1 million news ar- ticles. News articles have been gathered from more than 2000 news sources by ComeToMyHead in more than 1 year of activity. 4 labels: World, Business, Sports, and Sci-tech	Classification (World, Business, Sports, Sci-tech)
Rotten Tomatoes	Rotten Tomatoes is a dataset containing 5,331 positive and 5,331 negative processed sentences from Rotten Tomatoes movie reviews. (Positive, Negative)	Binary
SST2	SST2 consists of 11,855 single sentences extracted from movie reviews. It was parsed with the Stanford parser and includes a total of 215,154 unique phrases from those parsed trees. (Positive, Negative)	Binary
Yelp Polarity	The Yelp reviews polarity dataset is constructed by considering stars 1 and 2 as negative, and 3 and 4 as positive. In total, there are 560,000 training samples and 38,000 testing samples. Negative polarity is class 1, and positive class 2. (1, 2)	Binary

H. Running the Attack

After finally deciding on the datasets to be used, models to be tested, and the attack configuration, the attack is then run and tested. For this research, we kept the number of examples to 100 which means the attack will target 100 examples in the dataset.

The 'attackargs' function sets the attackargs value, spec-ifying that the attack should be used for 100 samples, and disables the output model. Finally, it performs an attack on the data using the Attacker category, attempting to modify the input texts to reduce the model's classification accuracy.

Algorithm 1: BEACOMP Adversarial Attack
Aim: Adversarial attack framework to fool neural text classifier
Input: legitimate input X and its ground truth label Y, classifier $F(.)$, threshold δ, semantic similarity $S(.)$.
Output: Generated an adversarial sequence X_{adv}

1. Initialization: $X^* \Leftarrow X$
2. for X_i in x do
3. Compute score (X_i^*) according to Equation (5)
4. end for
5. $W_{Ordered} \Leftarrow Sort(X_1, X_2, X_3 \ldots\ldots\ldots\ldots X_n)$ in descending order
6. Remove the stop words in $W_{Ordered}$
7. for X_i in $W_{Ordered}$ do
8. $X^* \Leftarrow replacingXwith(newword \notin$ *(insertion + replacement)*$)$
9. if $S(X, X^*) <= \delta$ then
10. return None:
11. else if $F(X^*) \neq Y_{true}$, then
12. Solution found. return X^*.
13. end if
14. end for
15. return None

III. RESULTS

Before discussing, the results, let us get into what a Beacomp attack looks like. Shown in BEMOCOMP algorithm, is the snippet of one of the attacks made on the Yelp-polarity dataset using the BERT Uncased model in which it is clearly shown how many and what words are being changed in order to produce a successful adversarial input to the model. For instance, in the example given below, 4 words were changed from their original words through which, the attack was able to successfully fool the model and was able to hence, produce a successful adversarial attack. Through this, it can also be seen that the Average perturbed rate is not a very efficient way of discussing the efficiency of the attack as even with a low APR rate in this example, the attack was successful. However, APR is important in order to understand the overall working, the robustness, and to understand the efficiency of the model. This is the reason that we have kept APR as one of the parameters while evaluating the success rate of the attack.

TABLE II. ACCURACY OF THE MODEL BEFORE AND AFTER THE ATTACK (IN %)

Model	**Distilled BERT**		**BERT Uncased**		**RoBERTa**	
	Before	**After**	**Before**	**After**	**Before**	**After**

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 20,2024 at 06:19:33 UTC from IEEE Xplore. Restrictions apply.

AG News	99.0	1.0	98.0	2	100.0	0
Rotten Tomatoe	98.0	0	100.0	10	85.0	1.0
SST2	99.0	2.0	100.0	0	92.0	4.0
Yelp Polarity	99.0	1	100.0	0	96.0	2.0

TABLE III. ASR (ATTACK SUCCESS RATE) AND APR (AVERAGE PERTURBED RATE) OF THE ATTACK

Model	**Distilled BERT**		**BERT Uncased**		**RoBERT**	
	ASR	**APR**	**ASR**	**APR**	**ASR**	**APR**
AG News	98.99	37.36	100	58.11	97.92	14.24
Rotten Tomatoe	92.28	15.36	90	18.56	98.82	14.50
SST2	90.0	9.64	100	38.08	93.59	9.88
Yelp Polarity	96.62	32.34	100.0	4.01	99.26	11.42

The success of the attack was measured on three different parameters, namely the change in accuracy of the model before and after the attack, the attack success rate(APR), and the Average perturbed rate(APR). By seeing the results, it is pretty evident that the attack was able to significantly lower the accuracies of all three different models tested. Also, as one of the parameters taken into consideration while measuring the success of the attack, the attack success rate was seen to be at an average of 96.46%, which showcases a very high percentage of attacks being successful.

IV. CONCLUSION

With all the results of the attack discussed above shown in table number 2 and 3, it is pretty evident that the novel attack BEACOMP attack was successfully able to fool all the models with a very good efficiency, and Attack success rate. The attack was able to differentiate itself from the previously discussed attack algo- rithms by including a composite attack composition which also

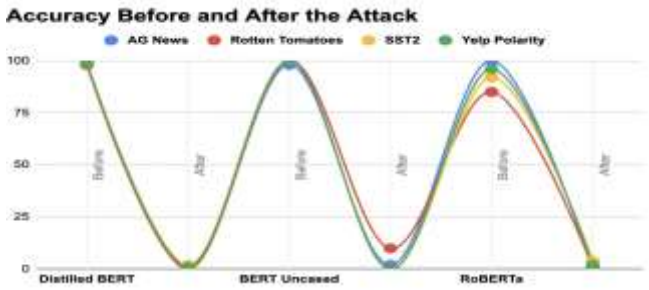


Fig. 8. Accuracy Before and After the attack

resulted in an increased efficiency of the attack on all three models. There was also a conclusion that the average perturbed word does not play a very large impact on the success of the attack. Ultimately, the attack was successful and on average, the model's accuracy dropped from 97.16% to 1.91%. This demonstrates the effectiveness of the TextAttack library in generating attack patterns for text classification patterns while adhering to certain constraints. This attack achieved a negative trend in altering inputs, highlighting the importance of trying to strengthen and prevent attacks such as attacks on natural language use.

V. DISCUSSION AND FUTURE WORK

The robustness and the accuracy of NLP models are be- coming a significant area of research which is vital to all the applications of NLP in today's world. With the increasing usage of NLP models, it is important that the vulnerability of these models are taken into consideration when working with them. In the future, many possibilities of research in the same area are needed and possible, some of the areas plausible are:

A. *Adversarial Attack Variants :* Different variants of attack recipes are possible through the textattack library which will help in discussing and comparing the robustness of NLP models on different attack compositions. The composition of different at- tacks will also help in creating a more robust attack composition for future research studies. New types of attacks can also be developed such as targetted attacks, where the attacker aims to make the model classify the input as a specific class, knowledge-based attacks, where the attacker has access to additional information about the model, data, or training process, or developing attacks for models other than binary outputs.

B. *Adversarial Defense Mechanisms:* Along with the dif- ferent variants of attack recipes, the defence mech- anisms of these models to these adversarial attacks also play a large part to play. Different mechanisms can be developed for models in order to successfully defend themselves from the adversarial attacks and will hence, increase the robustness, and will decrease the vulnerability of these models. Different techniques like including adversarial training in building the model, in- put preprocessing, and model architecture improvements can be implemented in order to increase the quality of the defence mechanisms of these models.

C. *Ethical and Security Considerations:* With the rapid increase in the usage of NLP models, the ethical and security considerations become a very vital part to take into consideration when dealing with them. While improving the defense mechanism, or introducing new attack variants, the privacy of the data incorporated in the data should not be hampered. Also, through these attacks, it should be taken into consideration that the personal data in the data should not be made vulnerable to cyber attacks, so as to prevent distrust in using NLP models.

D. *Adversarial Attacks on Neural Language Generation:* Generative AI is one of the fastest emerging fields currently in the field and it is very important to consider the possibility of applications made based on Generative AI to be robust to these adversarial attacks. Hence, robustness to adversarial attacks should be incorporated into training these applications. Some of the fields in which this can be done include text generation and machine translation, both of which are very important fields in Natural language processing and are used on a large scale by many companies and people around the world.

E. *Adversarial Attacks in Multilingual:* Settings More often than not, the idea of Multilingual attacks is left out while discussing new research ideas and as the majority of the population of the world does not have English as their primary language, it is very important to include different geographies and languages while discussing and developing these attacks.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 20,2024 at 06:19:33 UTC from IEEE Xplore. Restrictions apply.

For instance, machine translation is one of the most prominent fields applicable for using adversarial attacks in Multilingual settings.

REFERENCES

[1]. L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," Proceedings of the 4th ACM workshop on Security and artificial intelligence - AISec '11, 2011, doi: https://doi.org/10.1145/2046684.2046692.

[2]. Laskov and R. Lippmann, "Machine learning in adversarial environ- ments," Machine Learning, vol. 81, no. 2, pp. 115–119, Aug. 2010, doi: https://doi.org/10.1007/s10994-010-5207-6.

[3]. A. Bajaj and D. K. Vishwakarma, "HOMOCHAR: A novel adversarial attack framework for exposing the vulnerability of text based neural sentiment classifiers," Eng Appl Artif Intell, vol. 126, Nov. 2023, doi: 10.1016/j.engappai.2023.106815.

[4]. Bajaj and D. Kumar Vishwakarma, "Evading text based emotion detection mechanism via adversarial attacks," Neurocomputing, vol. 558,p. 126787, Nov. 2023, doi: 10.1016/j.neucom.2023.126787.

[5]. A. Bajaj and D. K. Vishwakarma, "Bypassing Deep Learning based Sentiment Analysis from Business Reviews," in 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), IEEE, May 2023, pp. 1–6. doi: 10.1109/ViTECoN58111.2023.10157098.

[6]. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment", AAAI, vol. 34, no. 05, pp. 8018-8025, Apr. 2020.

[7]. Sidorov, G., Gelbukh, A., Go´mez-Adorno, H., & Pinto, D. (2014, September 30). Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. Computacio´n Y Sistemas, 18(3). https://doi.org/10.13053/cys-18-3-2043

[8]. S. Lai, K. Liu, S. He and J. Zhao, "How to Generate a Good Word Embedding," in IEEE Intelligent Systems, vol. 31, no. 6, pp. 5-14, Nov.- Dec. 2016, doi: 10.1109/MIS.2016.45.

[9]. Jin Yong Yoo, J. X. Morris, E. Lifland, and Y. Qi, "Searching for a Search Method: Benchmarking Search Algorithms for Generating NLP Adversarial Examples," arXiv (Cornell University), Jan. 2020, doi: https://doi.org/10.18653/v1/2020.blackboxnlp-1.30.

[10]. J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP," arXiv:2005.05909 [cs], Oct. 2020, Available: https://arxiv.org/abs/2005.05909

[11]. Bajaj and D. K. Vishwakarma, "Exposing the Vulnerabilities of Deep Learning Models in News Classification," in 2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT), IEEE, Feb. 2023, pp. 1–5. doi: 10.1109/ICITIIT572462023.10068577.q

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 20,2024 at 06:19:33 UTC from IEEE Xplore. Restrictions apply.

https://doi.org/ 10.33472/AFJBS.6.9.2024.672-696

Bio Synergize: Microbial Synergy Driving Simultaneous Bioremediation and Nanoparticle Synthesis

Arif Khan[1], Suman Yadav[2], Jai Gopal Sharma*[3]

[1, 2,3]Department of Biotechnology, Delhi Technological University, Delhi-110042, India

*Correspondence:

Jai Gopal Sharma,

Department of Biotechnology, Delhi Technological University, Delhi-110042, India

Article History

Volume 6,Issue 9, 2024

Received:22 Mar 2024

Accepted : 27 Apr 2024

doi: 10.33472/AFJBS.6.9.2024.672-696

Abstract.

Heavy metals even in a very low quantity can be hazardous which can cause significant ecological and health consequences affecting the physiological functioning and structural makeup of all the living things. The elevated level of heavy metals in the environment causes biodiversity loss, environmental pollution, and ecosystem alteration. thus, it is imperative to remove heavy metals from contaminated region through detoxification. Using microorganisms such as fungi, bacteria, algae and yeast is a cost-effective, sustainable, and promising method in order to bioremediate heavy metals. Living organisms can alter contaminants as part of their metabolic processes, breaking them down into less toxic substances. The distinct advantage of microbial remediation is the rate of heavy metal cleanup, which is much faster than phytoremediation. Microorganisms absorb and detoxify heavy metals faster than plants, avoiding the spread of contaminants in the surrounding environment. The ability of microorganisms to execute the detoxification of harmful metal ions by reducing the metal ion into metal nanoparticles makes them appropriate for use in the simultaneous elimination of harmful chemicals from the polluted area and biosynthesis of nanoparticles which holds an immense potential to be used in biomedical and industrial applications. This review focuses the process of detoxifying heavy metals by using microorganisms to synthesize nanoparticles and remediate the affected areas simultaneously.

Keywords: heavy metals, bioremediation, bioaccumulation, synergy, nanoparticles biosynthesis

1. Introduction.

During the last hundred years, industrialization has grown at a tremendous rate. This also increased the requirement for the earth's natural resources; these resources are utilized at a careless rate, which leads to huge contamination of the earth's environment. The biggest contributors to environmental contamination are pesticides and the dumping of industrial waste into water bodies which leads to water pollution. The xenobiotic substances, which are produced by humans, are resistant to natural degradation thus, it's a matter of great concern because accumulation of the xenobiotic substances can lead to environmental pollution and can affect various life forms. These pollutants are classified as either organic or inorganic pollutants. Some examples of typical organic pollutants are azo dyes, petroleum hydrocarbons, chlorinated phenol, persistent organic pollutants (POPs), insecticides, etc. Heavy metals are inorganic pollutants, which include, lead (Pb), chromium (Cr), mercury (Hg), nickel (Ni), copper (Cu), cadmium (Cd), manganese (Mn), cobalt (Co), iron (Fe) (Pavan Kumar Gautam & Ravindra Kuma, 2022). Metallic elements that have large atomic weights or high atomic numbers, and high densities, are classified as heavy metals. These metallic elements have greater density than $5g/cm^3$ (Duruibe et al., 2007). Due to their occurrence in microscopic amounts, heavy metals are often referred to as trace elements. Heavy metals even in a very low quantity can be hazardous and can cause significant ecological and health consequences affecting the physiological functioning and structural makeup of all living things. The heavy metals are quite toxic to various life forms and thus, it is necessary to carry out the detoxification or elimination of the heavy metal from the polluted site. The Conventional physical and chemical method used for removing the pollutants from wastewater can lead to the generation of by-products and secondary pollutants and is thus not a cost-effective and sustainable method for the elimination of heavy metals from the polluted sites (Chibuike & Obiora, 2014). An economical and ecologically sustainable approach needs to be used for the detoxification of heavy metals. Bioremediation is a method that involves utilizing living things like microorganisms and plants, that can break down a broad spectrum of xenobiotics compounds and can convert contaminants into less hazardous forms (Vidali, 2001a). Living organisms can alter contaminants compounds as part of their metabolic processes, breaking them down into less toxic substances. Bioremediation involves phytoremediation which involves the use of plants for the purpose of remediation and micro-remediation involves the use of microorganisms involving fungi, bacteria, and algae. Phytoremediation, despite being a potentially effective and environmentally friendly approach, has several drawbacks and limitations. One of its limitations is that the process takes a lot of time that may require months or years to achieve a substantial reduction in contaminants (Vidali, 2001b). Another safe, effective, and substitute approach for heavy metal eradication is micro remediation. Micro remediation involves the use of fungi, bacteria, and algae for the elimination of heavy metals from the polluted site by converting harmful heavy metals into less harmful forms or immobilizing them making them less toxic(Zhang & Zhang, 2022). Microbes are the preferred choice for this task as they are convenient to handle, cultivate, and implement. Instead of planting lots of plants, microorganisms might be introduced directly to the polluted sites. This

eliminates extensive planting and maintenance work, making it practical for the urban environment. The biosynthesis of nanoparticles can be performed using fungi, bacteria, algae, and plants. The biosynthetic process involves the reduction of metal ions into metal nanoparticles. The compounds present in microbes act as stabilizing and reducing agents for the biogenic synthesis of nanoparticles. Due to the high exposure to hazardous chemicals and elevated concentrations of heavy metals, the microorganism has developed some defense mechanisms to survive in such a harsh condition. This review focuses on the synergy of bioremediation of heavy metal by simultaneously synthesizing the nanoparticles using microbes. The microorganism has the ability to detoxify the harmful metal ions by either reducing the metal ion into zero valent state or metal nanoparticles making them suitable to be use for the simultaneous removal of harmful chemicals from wastewater and other contaminated sources and reducing the heavy metal ion into metal nanoparticles.

2. Sources of heavy metal

The earth's crust naturally contains heavy metals. Due to the dramatic growth in the usage of heavy metals, there is an increase in the concentration of metallic components in terrestrial and aquatic ecosystems. Anthropogenic activities including fossil fuel combustion, mining, metal leaching, sewage, oil refinery, lime, industrial waste, inorganic fertilizer, electronic waste, pesticides, and fungicides pesticides are the primary sources of heavy metal pollution (Yadav & Sharma, 2023). Higher quantities of Cr, Ni, Pb, and Cd can be found in inorganic fertilizers, especially in phosphate fertilizers. Cadmium is particularly a dangerous element for plants due to its higher degree of accumulation in leaves, which can be ingested by humans or animals. The excess of Cd causes kidney dysfunction, hypertension, bone fractures, and lung cancer in humans (Kumar Sharma & Agrawal, 2005). Natural sources of heavy metal pollution are soil erosion due to heavy rainfall, volcanic eruptions, comets, metal corrosion, decomposition, and geological weathering. When rocks, minerals, and soil establish a connection with water, atmosphere, and living things, a process known as weathering occurs that breaks them down. It is a site-specific process in which heavy metals are released from the host rock into the surrounding environment by the activity of microorganisms or micro-climatic factors including pressure, ice, water, and heat. (Wen et al., 2021) carried out a study and their findings revealed that weathering causes the release of heavy metals As, Hg, and Tl from the argillaceous sandstone. FTIR and Raman spectroscopy showed that prolonged weathering broke down carbonyl and hydroxyl functional groups, resulting in the release of bound heavy metals from the rocks (Wen et al., 2021). Metals and other substances are taken up as nutrition by the living things from their surroundings. These nutrients and metals are crucial for their development. These nutrients are returned to the environment after their death and decomposition. The most frequent heavy metals present in decomposing organic matter are Hg, Cd, Cr, Pb, As, and Zn (Stefanowicz et al., 2020). Volcanic eruptions are a natural process. It causes particles to be released onto the earth's surface. volcanic ash and pyroclastic material are emitted during eruptions of volcanoes. heavy metals are also found in volcanic ash because volcanoes are created by the metal dissolution from the earth's crust. Cd, Fe, Zn, Pb, Cu, and Cr can be found in volcanic ash (L, 2010). During rainfall,

they are frequently transmitted to the environment. Volcanic ash also contains volcanic dust, which is carried by the wind and lands on the ground when it rains. The trace metals that are present in volcanic dust go to the soil, and depending upon temperature, pH, and redox they dissolve in water and get absorbed by the soil particles (Ilyinskaya et al., 2021; Ma et al., 2019).

3. Approaches for the removal of heavy metals.

The elevated level of heavy metals in the environment causes biodiversity loss, environmental pollution, and ecosystem alteration. To keep the environment safer for living organisms, the high concentration of trace elements and heavy metals must be eliminated from polluted water bodies and soil. There are various chemical, biological, and physical methods that can be used for the purpose of remediation. It is possible to eliminate these heavy metals using a variety of techniques, reverse osmosis, membrane technology, oxidation, reduction, filtration, ion exchanges, membrane technology, electrochemical treatment, and evaporation. However, many of these strategies lose their effectiveness at heavy metals concentrations below 100 mg/L (Nourbakhsh et al., 1994). These Physio-chemicals methods are not ecologically friendly and relatively expensive for heavy metals remediation and have other drawbacks also such as inadequate metal removal rates, increased need for energy, and reagents, and the production of hazardous sludge. Additionally, these physio-chemical methods are not suitable for extensive contaminated sites such as industrial waste sites, agriculturally polluted soils, and mining areas due to their high costs.

Hence, ecologically friendly and inexpensive techniques are required for the remediation of metal-contaminated water and soil.

3.1. Biological approach for remediation.

A biological technique called bioremediation in which environmental contaminants are excluded with the assistance of living things including microorganisms and plants which can break down a broad spectrum of xenobiotics compounds and convert contaminants into less hazardous forms (Vidali, 2001a). Living organisms can alter contaminants compounds as part of their metabolic processes, breaking them down into less toxic substances.

Bioremediation has many advantages over physio-chemical techniques, they are less costly when compared to the conventional methods, and they offer a permanent approach that could result in the complete mineralization of contaminants. Physio-chemical methods have a major disadvantage in that they cannot remove contaminants when they are present in a very low concentration, but with the help of bioremediation low concentration contaminants can be treated effectively (Kour et al., 2021).

Phyto-remediation and micro-remediation are the two different terms for bioremediation, in phytoremediation an essential part is played by the plants in the remediation of contaminants. On the other hand, micro remediation is conducted by microorganisms such as fungi, algae, and bacteria.

3.1.1. Phytoremediation

The word “phytoremediation” was introduced in 1991, and is composed of the fusion of two words: Greek phyton (meaning plant) and Latin remediare (meaning remedy). Plants have

natural inherent capabilities to absorb, bioaccumulate, preserve, or decompose inorganic or organic chemical compounds (Ali et al., 2013). When compared to other remediation methods, phytoremediation has very low Maintenance and installation costs also applicable for enormous, contaminated area where other alternative remediation methods are neither practical nor economical (Van Aken, 2009). Furthermore, phytoremediation not only cleans up polluted environments but also provides benefits such as biofuel production and carbon sequestration (Doty et al., 2007). (Ranieri et al., 2016) used the plant species *Phragmites Australis* and *Ailanthus Altissima* and removed between 55 to 61% of Cr^{3+} from the water. In both plants, a significant amount of chromium is absorbed by the roots (Ranieri et al., 2016). For heavy metal phytoremediation, hyperaccumulators are most appropriate. Hyperaccumulators are the plant species that have higher efficiency for the metal's accumulation from the soil and storage in their tissues as compared to normal plants (Dinh et al., 2004). They can accumulate concentrations 100 times higher than those of regular plants growing on sites contaminated with metals (Liu et al., 2022). Over 400 species renowned for their exceptional ability to hyperaccumulate metals are from the families Brassicaceae, Asteraceae, Rubiaceae, and Euphobiaceae (McGrath & Zhao, 2003). Hyperaccumulator plants also demonstrate efficient metal transport from roots to shoots. The transport efficiency of non-hyperaccumulators from roots to shoot when compared with the efficiency of hyperaccumulators is less although they accumulate metals in their root in higher concentration. (Robinson et al., 2003) used the rapidly growing hyperaccumulator plant *Berkheya coddii* for the remediation of nickel. They grew the plants hydroponically, and the nutrient solution carries varying concentrations of Ni. To find the concentration of Ni, they employed inductively coupled plasma emission spectroscopy (ICP-AES) and found that the highest Ni concentration was in the leaves and accumulated in a water-soluble state. Ni concentration was greater in the shoot than in the roots (Robinson et al., 2003). Heavy metals accumulation efficiency of plants could be increased by supplying chelators such as EDTA, oxalic acid, and DTPA. Together with the metals in the contaminated soil, they form a complex and enhance the rate of transfer of metals to plant roots and above-ground components. (Ehsan et al., 2014)experimented to investigate, Citric acid effects on the phytoremediation of Cd in *Brassica napus L.* They grow the plants hydroponically and Two weeks later, $CdCl_2$ and citric acid were introduced into the nutrient solution. Plants were then taken out of the solution after 8 weeks. They used flame atomic absorption spectroscopy to measure the amount of Cd in the different parts of the plant and found that, in comparison to plants grown without citric acid, there was a high concentration of Cd in the plant's roots and shoot (Ehsan et al., 2014).

Phytoremediation, despite being a potentially effective and environmentally friendly approach, has several drawbacks and limitations. One of its limitations is that it is a long process that may require months or years to achieve a substantial reduction in contaminants. Also, the selection of suitable plants is important, as not all plants possess the ability to effectively absorb contaminants. It is appropriate for minimal to moderately heavy metal-contaminated regions, since extremely contaminated sites don't promote plant growth and there are concerns regarding the transfer of heavy metals to the food chain if not properly monitored (Vidali, 2001b).

(Bañuelos et al., 2002) carried out various studies, and their findings highlighted that the plants used in the phytoremediation of selenium could transfer Se to animals and insects unintentionally. Insects feeding on these plants have the capacity to accumulate large amounts of selenium. The accumulated high selenium could harm animals and birds that feed on these insects (Bañuelos et al., 2002).

Another safe, effective, and substitute approach for heavy metal eradication is micro remediation, which utilizes fungi, bacteria, and algae to convert harmful heavy metals into less harmful forms or immobilize them making them less toxic. Microbes are the preferred choice for this task as they are convenient to handle, cultivation, and implementation. Microbes are ubiquitous and can survive even in adverse environmental situations such as high pH, temperature, and salt concentration. The microbes are incredibly tiny; their ratio of surface area to volume is high, so their rate of absorption is very high (Zouboulis et al., 2004). They have a rapid multiplication rate and can even work in conditions when extremely low concentrations of heavy metals are found.

The distinct advantage of microbial remediation is the rate of heavy metal cleanup, which is much faster than phytoremediation. Microorganisms absorb and detoxify heavy metals faster than plants, avoiding the spread of contaminants in the surrounding environment. Micro remediation is ideal for cities or small areas to clean polluted water or soil. Instead of planting lots of plants, microorganisms might be introduced directly to the polluted sites. This eliminates extensive planting and maintenance work, making it practical for the urban environment.

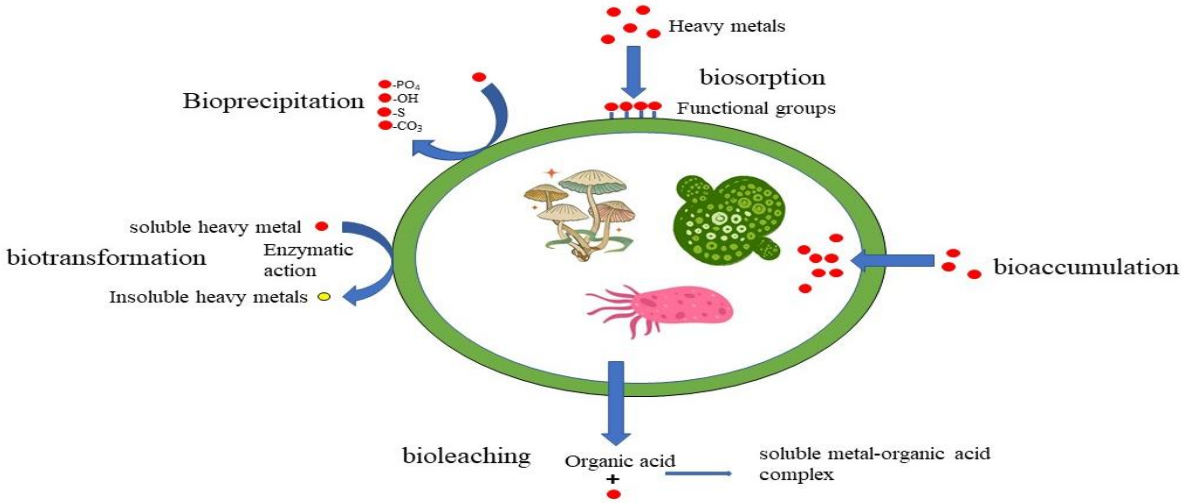


Figure 1. Mechanism of heavy metals remediation in fungi, algae, and bacteria.

3.1.2. Bacteria

Heavy metals also have a toxic effect on bacteria when they are growing in an environment containing high quantity of heavy metals beyond the threshold limit. Bacteria developed distinct mechanisms including biosorption, bioaccumulation, bio-precipitation, and bioleaching to protect themselves against toxic metals, and these metal-resistant bacteria can be utilized as effective bioremediation agents. The first line of defense against heavy metal toxicity is adsorption, in which metal ions bind nonspecifically to the surface of cells. Bacterial cell walls

have various types of functional groups, like phosphoryl, sulfate, amino, and carboxyl groups, with which heavy metals interact and get accumulated on the bacterial cell wall's surface. (Musa Jibrin et al., 2020) used *Lysinibacillus fusiformis* 5B for the biosorption of Cr, Ni, Pb, and Cd. Different concentrations of heavy metal-containing culture broths were prepared and inoculated with L. *fusiformis* 5B. For 35 days, inoculated culture broths are kept at 37C^{o} in a shaker incubator. Atomic absorption spectroscopy was used to quantify the amount of heavy metal biosorption, L. *fusiformis* 5B effectively removed heavy metals from the culture broths and can be used as biosorption agents (Musa Jibrin et al., 2020). Biosorption was observed in both live and dead bacteria cells. dead microbial biomass could be used for bioremediation. (Mohapatra et al., 2019) carried out a study, that used high Pb (II) concentration resistance *Bacillus xiamenensis* PbRPSD202 and analyzed biosorption capacity. Dead and active biomass of bacteria is employed for the elimination of Pb (II) from the aqueous solution. They used the XRD and FTIR techniques to investigate the function of ligands like, hydroxyl, amino, carboxyl, and carbonyl groups present on the bacterial surface in Pb (II) biosorption. Under optimum situations, dead and active biomass of bacteria removed 97.18% and 99.19% lead (II) from the solution (Mohapatra et al., 2019). Utilization of dead biomass of bacteria for bioremediation is becoming more prevalent as no nutrition is required for the bacterial cell's growth, thus reducing the cost of operation. Many bacteria produce extracellular polymeric substances (EPS) that have a major role in the production of biofilms and the adherence of bacteria to different surfaces. The essential components of these substances are polysaccharides, lipids, nucleic acids, and proteins. EPS are engaged in the defense of bacteria against toxic heavy metals, as they consist of various types of anionic and cationic functional groups. These functional groups on EPS interact with the heavy metals, preventing their entry into the bacterial cell, and accumulating them. (Cui et al., 2020) extracted the EPS from *Agrobacterium tumefaciens* F2 that showed effective absorption efficiency for the three heavy metals. Absorption efficiency of Ni^{2+}, Pb^{2+}, Cd^{2+} on EPS was found to be 77.95%, 94.67%, and 94.41% respectively. FTIR analysis showed that polysaccharides are the primary component in the EPS extracted from *Agrobacterium tumefaciens* F2 and have a crucial role in the absorption of Cd^{2+}, Ni^{2+}, and Pb^{2+} ions (Cui et al., 2020). Bacterial cells could alter the ionic state of heavy metals after adsorption, hence lowering their toxicity. Another bioremediation mechanism observed in bacteria is bioaccumulation in which heavy metals are accumulated inside the bacterial cells. When compared to biosorption, bioaccumulation is a more time-consuming and energy-dependent process. (Ahemad & Malik, 2011) isolated 34 bacteria from the heavy metal-contaminated soil, from which pseudomonas isolate SN30, pseudomonas isolate SN7, and pseudomonas isolate SN28 were the zinc-resistant bacteria and shown resistant against Cr^{3+}, Cr^{6}, Hg^{2+}, Cu^{2+} Ni^{2+}, Pb^{2+}, and Cd^{2+}. These bacterial isolates have shown the ability for the accumulation of zinc and copper. When the concentration of copper in the medium is 2.92 mM, bacterial isolates SN30, SN7, and SN28 accumulated 22, 20, and 25 mg/g dry weight of cells. Similarly, when the concentration of zinc is 1.6 mM, zinc resistance bacteria SN30, SN7, and SN28 accumulated 26, 29, and 25mg/g dry weight of cells (Ahemad & Malik, 2011).

3.1.3. Fungi

Fungi are an excellent option for the biological remediation of heavy metals, as they exhibit high-level metal tolerance and impressive wall and intracellular binding of metals. compared to other microorganisms, they are easier to grow, high biomass producers and low-cost media can be used for their growth (White et al., 1995). The Fungal cell wall usually consists of chitin, proteins, polysaccharides (80- 90%), lipids, inorganic ions, and phosphates. These components of fungal cell walls play a crucial part in heavy metal adsorption. There are various negatively charged functional groups like carboxyl, phosphate, amino, sulphate, hydroxyl, and amino groups present on the fungal cell wall. During adsorption, these functional groups bind with various metal cations. (Narolkar et al., 2022) isolated 9 fungal isolates from the metal-polluted site. At 150 ppm of Cr (VI), fungal isolate *Aspergillus candidus* demonstrated an impressive potential for adsorption of 5.49 mg/g and a chromium removal efficacy of 98.75% (Narolkar et al., 2022). (Kurniati et al., 2014) isolated filamentous fungi from the forest soil and investigated the potential of filamentous fungi for mercury bioremediation. *Aspergillus flavus* strain KRP1 exhibited high mercury resistance and demonstrated tolerance up to 100 mg/L. Mercury (II) was eliminated from the PBD media by strain KRP1 via biosorption. After incubating for 7 days, the Removal efficiency of KRP1 in static and shaken conditions is 98.73% and 97.50% in the PBD media containing 10 mg/L mercury (II) (Kurniati et al., 2014). Dead and live fungal biomass could be used as effective bio-sorbents because both have a variety of functional groups on their cell walls which can effectively used for getting rid of harmful heavy metals. (Ranjani, 2019) carried out a study and evaluated the biosorption ability of dead and live biomass of *Rhizopus stolinifer* and *Aspergillus niger* for chromium (VI), isolated from the metal-polluted leather tannery location. *Aspergillus niger* has demonstrated tolerance for chromium (VI) up to 100 ppm and Rhizopus *stolinifer* had a tolerance of up to 1000 ppm. The removal ability of dead and live biomass of *Aspergillus niger* was 45% and 25% and that of Rhizopus *stolinifer* was 83% and 45% at 100 ppm of chromium (VI) (Ranjani, 2019). Fungi exhibit another process by which heavy metals are bioremediated i.e bioaccumulation. It deals with the movement of metals through the cell membrane and accumulation inside the fungal cell. (Dursun et al., 2003) studied the bioaccumulation capability of *Aspergillus niger* and the effect of pH on the growth and bioaccumulation of lead (II), chromium (VI), and copper (II). The ideal pH for the bioaccumulation of lead (II), chromium (VI), and copper (II) and the growth of *Aspergillus niger* were 4.5, 3.5, and 5.0. At concentrations of 100 mg dm^{-3} of lead (II) and copper (II), the maximum accumulation capability of *Aspergillus niger* was 34.4 and 15.6 mg/g, respectively. It was observed that an increase in the concentration of chromium (VI) from 50 to 75, there is no fungal growth, and chromium (VI) uptake occurs. The maximum accumulation capability of *Aspergillus niger* at 50 mg dm^{-3} was 6.6 mg/g (Dursun et al., 2003).

3.1.4. Algae

Algae are quite effective in eliminating heavy metals from polluted environments. The majority of heavy metals are eliminated by algae via biosorption, accumulation, or biotransformation. On the surface of Algal cells, there are a variety of functional moieties present, metal ions bind with

these functional groups and get adsorbed at the cell wall. Biosorption is a passive process; dead and living algal biomass can be used as biosorbent. (El-Naggar et al., 2018) used the dried biomass of *Gelidium amansii* for the eradication of lead from the aqueous solution. The FTIR investigation indicated the presence of phosphate, carbonate, methylene, phenolic, and carbonyl groups on the cell wall; these functional groups have a key role in the biosorption of lead. Under optimal conditions, *Gelidium amansii* demonstrated 100% lead removal efficiency from the aqueous solution (El-Naggar et al., 2018). (Shamshad et al., 2016) evaluated the biosorption capacity of *Oedogonium westti* for Cr, Ni, Pb, and Cd. *Oedogonium westti* was cultured in a selected heavy metal-containing aqueous solution for 7 days. The findings of this investigation showed that the adsorption capacity of *Oedogonium westti* for Cr, Ni, Pb, and Cd was 0.620, 0.418, 0.261, and 0.974 mg/g, respectively. Concentrations of metal and pH greatly affect the removal capacity. When pH was 4 and Cr and Cd were present in low concentrations, the removal capacity of *O. westti* was 93% and 95%, respectively. In the case of Ni, increasing the amount of Ni in the aqueous solution also enhanced the removal capacity and reached 87%. The removal capacity for Pb increases with a rise in the pH and metal concentration; at pH 6, the removal capacity for Pb was between 91-96% (Shamshad et al., 2016). Biosorption inhibits the penetration of heavy metals in the algal cell, but when extremely high concentrations of heavy metals exist, they can enter the algal cell. Then algal cells exhibit another mechanism known as bioaccumulation, in which heavy metals are accumulated in the algal cell to mitigate their toxic effects. Bioaccumulation is an energy-dependent and slow process, only living cells can carry out this mechanism. (Pham, 2019) Carried out a study, and their findings revealed that *Scenedesmus sp.* is Able to bioaccumulate Cu. when the amount of Cu in the solution is 2 mg/l, Scenedesmus sp. demonstrated 89.5% Cu removal efficiency within 7 days (Pham, 2019). Similar findings were provided by (Vigara Fernández et al., 2021) that *Chlorella sorokiniana* accumulated 11232 mg/kg of Cd, and demonstrated 65% removal efficiency (Vigara Fernández et al., 2021).

Table 1. Microbial mediated bioremediation of heavy metals.

Type of microorganism used	**Heavy metal to be removed**	**Mechanism of bioremediation used to detoxify the heavy metal**	**Performance of bioremediation**	**Time taken for bioremediation**	**Reference**
Bacteria:					
Arthrobacter viscosus	Cr	Biosorption.	100%	144h	(Hlihor et al., 2017)
EB L14	Cd, Cu, Pb.	Bioaccumulation, biosorption.	75.78%,21.25%, 80.48%.	24h	(Guo et al., 2010)

Bacillus cereus	Hg	Biosorption.	104.4 mg/g.	72h	(Sinha et al., 2012)
Bacillus anthracis PS2010	Co, Zn, Pb, Cd, Cu	bioaccumulation	4.75, 5.22, 6.44, 3.41, 2.03 mg/g.	-	(Sobhy et al., 2014)
Acidithiobacillus ferrooxidans	Fe, Zn, Cu, Pb	Bioleaching	85.42%, 98.73%, 96.44%, 88.90%	28 days	(Rouchalova et al., 2020)
Micrococcus luteus DE2008	Cu, Pb	Biosorption.	408mg/g, 1965mg/g.	12h	(Puyen et al., 2012)
Lactobacillus plantarum MF042018	Cr, Ni	Bioaccumulation	30.2 %, 33.8%	24h	(Ameen et al., 2020)
Variovorax boronicumulans	Zn, Pb, Cd	Bioprecipitation	73.81%, 95.93, 73.81%.	72h	(Jalilvand et al., 2020)
Stenotrophomonas rhizophila	Zn, Pb, Cd	Bioprecipitation	63.91%, 96.25%, 71.3%	72h	(Jalilvand et al., 2020)
Sporosarcina pasteurii	Zn, Pb, Cd	Bioprecipitation	94.83%,98.71%, 97.15%	72h	(Jalilvand et al., 2020)
Pseudomonas fluorescens	Cd	Biosorption	99%	5 min	(Sankarammal et al., 2014)
Oceanobacillus profundus	Pb, Zn	Biosorption	94%, 54%	24h	(Mwandira et al., 2020)
Fungi					

Trichoderma fungus	Cd	Biosorption	78.46%	120 min	(Bazrafshan et al., 2016)
Pleurotus ostreatus	Zn, Ni, Cr, Cu	Biosorption	3.22, 20.40, 10.75, 8.06 mg/g	150 min	(Javaid et al., 2011)
Aspergillus lentulus FJ172995	Ni, Pb, Cu, Cr	Bioaccumulation	42%, 100%, 78%,79%	5 days	(Mishra & Malik, 2012)
Alternaria alternata	Ag, Hg, Cu, Pb	Bioaccumulation	63.1%, 68.5%, 76.4%, 80%	7 days	(Verma et al., 2016)
Komagataella phaffi	Cu, Pb, Cd	Bioaccumulation	21.63, 20.63, 25.23 mg/g.	7 days	(Liaquat et al., 2020)
Fusarium sp. MMT1	Cr	Bioaccumulation ,Biotransformation	100%	72h	(Guria et al., 2014)
Neurospora crassa	Cu, Pb	Biosorption	12.28,49.06 mg/g	15 min	(Kiran et al., 2005)
Neosartorya fischeri	Tl	Biosorption, Bioaccumulation	62.01, 432.91 mg/kg	30 days	(Urík et al., 2010)
Trichoderma gamsii	Cr	Biosorption	89%	420 min	(Kavita & Keharia, 2012)
Algae					
Porphyra leucosticta	Pb, Cd	Biosorption	95%, 75%	120 min	(Ye et al., 2015)
Scenedesmus sp.	Pb	Bioaccumulation	83.5-84.2%	7 days	(Pham et al., 2020)
Spirulina sp.	Mn, Se, Zn,		99.73%,98.83%,		(Mane & Bhosle, 2012)

Spirogyra sp.	Fe, Cr, Cu	Biosorption	79%, 98.93%, 98.3%, 81.2%	7 days	(Mane & Bhosle, 2012)
Botryococusbrur auni	Mn, Se, Zn, Fe, Cr, Cu	Biosorption	99.6%, 98.16%, 81.53%, 99.73%, 98.23%, 89.6%	7 days	(Uddin & Lall, 2019)
Spirulina platensis	Cu, Pb, Cd	bioaccumulation	82%, 93%, 89%	90 min	(Al-Homaidan et al., 2014)
Desmodesmus sp.	Cu	Biosorption	90.6%	90 min	(Rugnini L et al., 2018)
Scenedesmus obliquus	Ni, Cu	Biosorption	85%, 94%.	2 days	(Rinanti et al., 2018)
Phacus genera	Cu	Biosorption, Bioaccumulation	65.54% ± 8.04	240 min	(Ahmad et al., 2020)
Spirulina (Arthrospira) platensis	Al, Pb, Ni	Bioaccumulation	64.28%, 79.17%, 66.67%	7 days	(Arunakumara et al., 2008)
Chlorococcum sp.	Pb	Bioaccumulation	188 mg/g	2 days	(Upadhyay et al., 2022)
	As	Bioaccumulation	322.43±2.95µg/g	10 days	

4. Simultaneous bioremediation and nanoparticle synthesis; synergy

Due to an increased exposure to hazardous chemicals and the elevated levels of heavy metals, certain defence mechanisms have been developed by microorganisms in order to survive in such harsh conditions, such as intracellular bioaccumulation, extracellular precipitation, and changes in efflux pumps. The microorganism has the ability to detoxify the harmful metal ions by either reducing the metal ion into a zero-valent state or metal nanoparticles, making them suitable for the simultaneous removal of harmful chemicals from wastewater and other contaminated sources and the biosynthesis of nanoparticles that have a potential to be used in different biomedical applications (Jadoun et al., 2022). The biosynthesis of nanoparticles can be executed using fungi, bacteria, algae, and plants. The microorganism provides a benefit over plant-mediated synthesis of nanoparticles i.e. they can reproduce easily when compared with plants. Microbes in the

presence of metal oxides can synthesize metal nanoparticles as well as they can be used for the biological remediation of heavy metals this synergy is due to the capability of microorganisms to tolerate the high concentration of heavy metals because of the presence of intracellular and extracellular proteins that mediate the adsorption and chelation of metal (Gomaa, 2022). The synthesis of metallic nanoparticles can occur either within the microbial cell or using the supernatant of the microbial cell which consists of different proteins, enzymes, and polysaccharides of the microorganism that act as reducing agents and are in charge for converting metal ions into metal nanoparticles.

The intracellular synthesis of nanoparticles involves three steps; trapping of metal ions onto the cell wall of microorganisms, bio-reduction by enzymes present in the cytoplasm, and capping and stabilization by organic molecules within the cell (Fariq et al., 2017). (Salvadori et al., 2014) using the dead biomass *of Rhodotorula mucilaginosa* carried out the intracellular synthesis of Cu nanoparticles. The copper ions were reduced by the dead biomass of red yeast into copper nanoparticles and apart from acting as a reducing agent the dead biomass also contributed in stabilizing the formed copper nanoparticles. The average diameter of the synthesized copper nanoparticles was observed to be 10.5nm and hence they concluded that the dead yeast biomass can be successfully used for the biosynthesis of copper nanoparticles with subsequent bioremediation of wastewater by removing copper from wastewater (Salvadori et al., 2014).

Some enzymes such as NADH-dependent enzymes which are released into the culture medium as extracellular enzymes serve in the metal ion reduction. The proteins, peptides, and other organic molecules found in the supernatant which is released by the microbial cell serve as capping and stabilizing agent for the synthesized nanoparticles (Murali et al., 2023). (Tyagi et al., 2019) used an entomopathogenic fungus, *Beauveria bassiana* for the biosynthesis of Ag nanoparticles. They used the extracellular synthesis approach for manufacturing Ag nanoparticles. The Supernatant was separated from the fungal biomass and in the supernatant silver nitrate was added which was further reduced to Ag nanoparticles. The color change of the solution to dark brown from white color was the first indication that confirmed the synthesis of Ag nanoparticles, further analysis of nanoparticles revealed that their size was in a range of 10-50nm and the Ag nanoparticles exhibit antimicrobial efficacy against different bacterial strains(Tyagi et al., 2019). (Ramanathan et al., 2013) used silver-resistant bacterium, *Morganella morganii* for the biosynthesis of copper nanoparticles and they evaluated the link between the biosynthesis of copper nanoparticles and the resistance that bacteria exert against heavy metal. Bacteria synthesize metalloproteins by utilizing copper, thus copper is a trace element that is utilized by bacteria in several ways however, at high concentrations, Cu (II) is toxic for the cell, and it can induce hydroxyl radical generation which can damage the cell in a different manner and can lead to microbial cell toxicity (Brow et al., 1992). Hence it is necessary for the bacterium to protect itself from the toxic concentration of copper ions and simultaneously must regulate the supply of copper to be utilized for the synthesis of metalloprotein. Further, their results indicated that the copper ions penetrate within the bacterial cell which within the cell associates with the biomolecules and is reduced to copper nanoparticles which are released out of the cells into the

medium. Their result also suggested that the Ag^{+} ions binding protein, SilE was involved in the uptake of Cu^{2+} ions inside the bacterial cell which further reduced into zero-valent Cu nanoparticles(Ramanathan et al., 2013). Bacteria, fungi, and algae can be effectively used in the bioremediation of heavy metals as well as in the synthesis of metal nanoparticles. Nitrate and silver are the common pollutants that are found in waste from ink, porcelain, and photographic manufacturing industries. (Eltarahony et al., 2016) used *Achromobacter sp.* MMT, which is a bacterial strain for the bioremediation of silver and nitrate pollutants, and due to the presence of periplasmic nitrate reductase enzyme the silver ions were simultaneously reduced to silver nanoparticles. For this purpose, they immobilized the bacteria and nitrate reductase enzyme so that they could withstand the harsh environmental condition of wastewater that was artificially made which contained $AgNO_3$. They observed that the immobilized bacterial cells were found to completely eliminate nitrate within 192 hours and simultaneously reduced Ag^{+} ions into Ag nanoparticles that had a size range between 23.26nm and 58.14 nm. The immobilized nitrate reductase enzyme had a lower efficiency in removing the nitrate from wastewater and the synthesized Ag nanoparticles by the enzyme were having a size range between 94.44 nm-172.22 nm(Eltarahony et al., 2016). (Sundararaju et al., 2020) used *Microbacterium sp.* for the removal of heavy metals from the surroundings, *Microbacterium sp.* is a bacterium that exhibits resistance to heavy metal and this bacterium was isolated from electroplating industrial effluent. Cobalt is a heavy metal that is released by some of the industries to the water bodies which can lead to a significant threat to the environment. The bacterium has shown resistance to Co (II) ions and is thus employed for the reduction of Co (II) ions to Cobalt oxide nanoparticles. Co_3O_4 nanoparticles were synthesized extracellularly using the *Microbacterium sp.* The Co_3O_4 nanoparticles were further characterized using different techniques and their analysis revealed their spherical morphology and the size range was observed to be 10–70 nm. These biosynthesized cobalt oxide nanoparticles can be utilized in the field of various industrial and clinical applications(Sundararaju et al., 2020). Phenol is an organic aromatic compound that at high concentrations is a toxic, carcinogenic, mutagenic compound it is released by pharmaceutic, petroleum, and other industries as a byproduct to the waterbodies and it can't degrade naturally. Thus, an ecologically friendly and economical method is employed for the bioremediation of phenol and tellurite oxyanions by (Hosseini, Lashani, et al., 2023). They used a bacterium, *Lysinibacillus sp.* which has an ability for the bioremediation of phenol and Te oxyanions. The bacteria utilized phenol as a source of carbon and reduced Te oxyanions to Te nanoparticles. The formation of Te nanoparticles was executed using an intracellular synthesis approach where the synthesized Te nanoparticles accumulate within the bacterial cell which was later on separated from the bacteria. The bacterium successfully degraded phenol by using it as a source of carbon and reduced tellurite oxyanions to Te nanoparticles and the formation of a black color of biomass indicated the formation of Te nanoparticles. The nanoparticles characterization indicated that the nanoparticles were observed to be spherical and their size range was observed to be 22 nm-148 nm. Thus, their study concluded that this biological method of bioremediation of these toxic

compounds is an effective mechanism for simultaneous bioremediation and nanoparticle synthesis(Hosseini, Lashani, et al., 2023) .

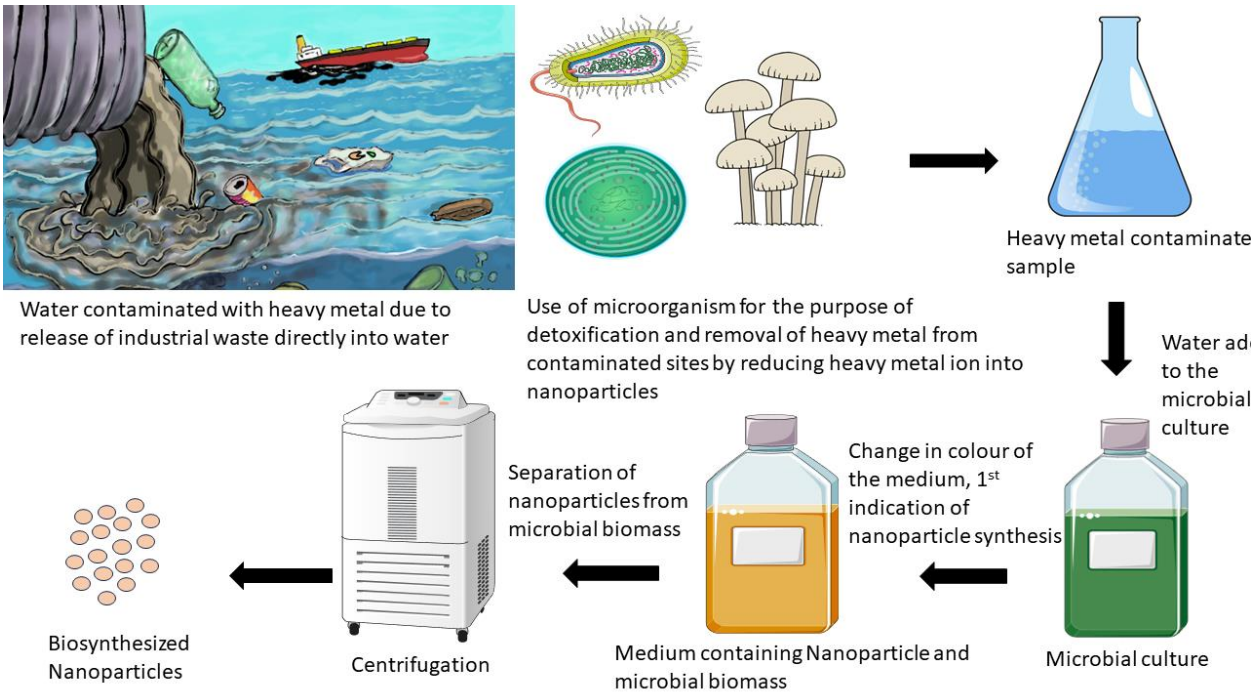


Figure 2. Schematic diagram illustrating simultaneous bioremediation of heavy metals from the contaminated sites and nanoparticle synthesis by reducing the heavy metal ion into nanoparticles

The presence of Cd in water and soil can lead to a significant danger to human life as it is a potent human carcinogen. Cadmium is a heavy metal and is quite toxic and thus it should be removed from the environment in order to prevent its exposure to humans through food and water. (Chakraborty et al., 2018) utilized extracellular polymeric substances released by *Pseudomonas aeruginosa* which is a bacterium isolated from marine areas for the bio-reduction of Cd^{2+}ions to CdS nanoparticles. The Extracellular polymeric substance released by *Pseudomonas aeruginosa* consists of polysaccharides, nucleic acids, proteins, and lipids which are accountable for the reduction of Cd^{2+} ions into CdS nanoparticles. The CdS nanoparticles were further analyzed using various techniques and they were characterized to be spherical and their size was observed to be around 10nm. They concluded that EPS secreted by *Pseudomonas aeruginosa* serves to be a sustainable method for the simultaneous removal of cd ions from the sample and the reduction of these cadmium ions to cadmium sulphide nanoparticles (Chakraborty et al., 2018). (Annamalai et al., 2013) used a biological method for the removal of chromium ions using *Bacillus subtilis.* At first, the biosorption of Cr (VI) was done to Cr (III) and was subsequently reduced to Cr nanoparticles using *Bacillus subtilis*. The manufacturing of Cr nanoparticles was done using an intracellular synthesis approach in which the Cr ions penetrate within the bacterial cell and subsequently inside the cell get reduced to Cr nanoparticles. Their results indicated that at pH 7 maximum chromium ions were removed by

Bacillus subtilis, further, the Cr nanoparticles were characterized to be spherical and their size ranges from 50 nm-78 nm. Thus, they provide sustainable and alternative approach for the detoxifying chromium ions from industrial effluents (Annamalai et al., 2013). Mercury is a heavy metal whose accumulation in the environment can lead to major problems in various life forms. Numerous human activities result in the release of Hg2+ ions into the environment such as the burning of coal, some industrial operations, and mining. Detoxification of Hg^{2+} ions can be done by carrying out the reduction of Hg^{2+} ions into HgS nanoparticles using various microorganisms. (Sathyavathi et al., 2013) used *Bacillus cereus* MRS-1 for the bioremediation of mercury by reducing Hg^{2+} ions into HgS nanoparticles. The major derivative of mercury is $HgCl_2$ which is released into the environment via various human activities. The bacterium has the ability to detoxify mercury chloride into HgS nanoparticles. In this study, $HgCl_2$ was added to *Bacillus cereus* MRS-1 culture, and the HgS nanoparticles were extracellularly precipitated. The biosynthesized HgS nanoparticles were observed to be spherical and their size ranges from 10 nm-100 nm. Thus, the *Bacillus cereus* strain which was isolated from industrial effluent of electroplating can be successfully used for the bioremediation of Hg which is present in the contaminated sites(Sathyavathi et al., 2013). (Chandrakant Vaigankar et al., 2020) carried out the detoxification of Selenite, using the bacterium *Halomonas venusta* which is a salt-tolerant bacterium. The bioremediation of selenite was done by reducing selenite to selenium nanoparticles by the salt-tolerant bacterium. The bacteria carried out the reduction of selenite oxyanions intracellularly and the fabrication of Selenium nanoparticles was first indicated when the color of the medium was changed to brick red from yellow. it was observed that the nanoparticles were spherical in shape and their size was observed in the range of 20 nm-80 nm also hexagonal crystal lattice was exhibited by them. They also evaluated anti-cancer, anti-oxidant, and also their potential against larvicidal mosquitoes (Chandrakant Vaigankar et al., 2020). Selenate and Selenite which are the oxyanions of selenium, these oxyanions are quite toxic at a high concentration to the aquatic and terrestrial life forms. However, some bacteria can withstand a high concentration of selenium ions and can reduce the oxyanions of selenium into selenium nanoparticles.(Pandey et al., 2021) carried out the biosynthesis of selenium nanoparticles in order to eliminate the toxic effect of selenium ions upon various life forms. They cultured *Anabaena sp.* For the biosynthesis of selenium nanoparticles. Sodium selenite was added to the cyanobacterial biomass and the solution was incubated. The nanoparticles were separated as supernatant and cyanobacterial biomass as pellets. Further, the selenium nanoparticles were characterized using different techniques, and the average diameter of the synthesized selenium nanoparticles was observed to be 25nm. The biosynthesized Se nanoparticles exhibited potent anti-bacterial activity and can be employed in the textile industry in the decolorization of dye (Pandey et al., 2021).

(Bedi et al., 2018) carried out the biosynthesis of iron-containing metal nanoparticles via bioremediation of mining waste. A significant amount of the mining waste is produced by the iron ore. They isolated a fungal strain from the mining site containing iron ore tailings *Aspergillus aculeatus*, which able to withstand elevated metal ion concentrations without leading

to its cell toxicity and can effectively convert the iron ore into metal nanoparticles. Their results indicated that the iron nanoparticles were highly effective in seed germination as these iron nanoparticles can be uptake by plants as a source of iron. Also, at low concentrations, they positively enhanced root and shoot development, however, they exerted toxic effects at higher concentrations on seedling growth (Bedi et al., 2018). Copper is the most common pollutant found in municipal wastewater, the presence of high concentrations of cupric ions can cause severe toxicity to various life forms. Conventional chemical and physical methods of removing pollutants from wastewater can lead to the generation of by-products and secondary pollutants. (Žvab et al., 2021) utilized the ability of *Chlamydomonas reinhardtii*, a microalga, for the bioremediation of nutrient media which is considered wastewater rich in Cu^{2+} ions. *C. reinhardtii* is used for the bioremediation of cupric ions by simultaneously biosynthesizing Cu nanoparticles. Further characterization of nanoparticles was done and the synthesized nanoparticles were observed to be spherical, and polydisperse and their size was observed to be 10nm. Thus, they effectively carried out the bioremediation of the Cu^{2+} polluted media and evaluated the potential of *Chlamydomonas reinhardtii* for the detoxification of metal ions by reducing it to metal nanoparticles (Žvab et al., 2021). Some natural activities such as the industrial activity, mining, burning of coal, and weathering of rocks can lead to the development of selenite and tellurite oxyanions which are quite toxic for living organisms. Thus, these oxyanions need to be removed from the environment. For this purpose, (Hosseini, Hadian, et al., 2023) carried out the bioremediation of selenite and tellurite oxyanions using *Trichosporon cutaneum* and *Yarrowia lipolytica*, these yeast strains reduced Se and Te ions into Se-Te nanoparticles. The nanoparticles were biosynthesized intracellularly by the yeast. The size of the Se-Te nanoparticles synthesized by *Yarrowia lipolytica* was found in a range of 46 nm-171 nm while the size of nanoparticles biosynthesized by *Trichosporon cutaneum* was observed in a range of 25 nm-53 nm. The analysis of synthesized nanoparticles revealed that the synthesized nanoparticles were covered by carbohydrates, proteins, and lipids and thus they successfully removed the oxyanions from the polluted environment by simultaneously synthesizing Se-Te nanoparticles which can be used in biomedical applications (Hosseini, Hadian, et al., 2023).

5. Future prospect

Research needs to be conducted in order to identify what can be done to improve the rate of biosynthesis of nanoparticles and increase the efficiency of microorganisms to more effectively carry out the bioremediation and detoxification of heavy metals. Additionally, some microorganisms used for simultaneous bioremediation and nanoparticle production could be pathogenic, so they could not be widely employed. Genetic engineering can be used to overcome these challenges. Modification in the genome of organisms can be done to alter the genes, and microorganisms with enhanced bioremediation capacities and less potential to cause diseases can be produced. The majority of these heavy metal remediation techniques are limited to the laboratory scale, employment of simultaneous bioremediation and nanoparticle synthesis at large a scale would be a major challenge. Apart from the biosynthesized nanoparticles have several benefits in different areas however, the nanoparticles can serve to some major health

complications if accumulated inside the human body in a high concentration. Thus, apart from the positive applications of nanoparticles the negative impact of nanoparticles on living organisms must also be evaluated.

6. Conclusion

The use of microorganisms such as fungi, yeast, bacteria, and algae is an economical, sustainable, and promising method for the bioremediation, detoxification, or elimination of heavy metals such as copper, cobalt, chromium, tellurite, selenite, cadmium, mercury and other contaminants from the contaminated sites in the environment. Some natural and anthropogenic activities cause these highly toxic heavy metals to release into the surroundings which could pose a significant danger to various life forms, thus it is essential to remove heavy metals from contaminated areas such as wastewater. The use of microorganisms has a dual benefit, they can effectively remove heavy metal from the surroundings by reducing it to nanoparticles, and the biosynthesized nanoparticles due to their smaller size can be employed in various biomedical applications such as antibacterial, anti-oxidant, anti-cancer, etc. thus this dual benefit can largely benefit the society in the removal of contaminants from the contaminated areas and also in the field of medicine, nanotechnology ad industrial applications.

Acknowledgment.

The author is very thankful to Delhi Technological University for providing infrastructure and all facilities.

References

Ahemad, M., & Malik, A. (2011). Bioaccumulation of Heavy Metals by Zinc Resistant Bacteria Isolated from Agricultural Soils Irrigated with Wastewater. *Bacteriology Journal*, *2*(1), 12–21. https://doi.org/10.3923/bj.2012.12.21

Ahmad, N., Mounsef, J. R., Tayeh, J. A., & Lteif, R. (2020). Bioremediation of Ni, Al and Pb by the living cells of a resistant strain of microalga. *Water Science and Technology*, *82*(5), 851–860. https://doi.org/10.2166/wst.2020.381

Al-Homaidan, A. A., Al-Houri, H. J., Al-Hazzani, A. A., Elgaaly, G., & Moubayed, N. M. S. (2014). Biosorption of copper ions from aqueous solutions by Spirulina platensis biomass. *Arabian Journal of Chemistry*, *7*(1), 57–62. https://doi.org/10.1016/j.arabjc.2013.05.022

Ali, H., Khan, E., & Sajad, M. A. (2013). Phytoremediation of heavy metals-Concepts and applications. In *Chemosphere* (Vol. 91, Issue 7, pp. 869–881). Elsevier Ltd. https://doi.org/10.1016/j.chemosphere.2013.01.075

Ameen, F. A., Hamdan, A. M., & El-Naggar, M. Y. (2020). Assessment of the heavy metal bioremediation efficiency of the novel marine lactic acid bacterium, Lactobacillus plantarum MF042018. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-019-57210-3

Annamalai, K., Mohan Nair, A., Chinnaraju, S., & Kuppusamy, S. (2013). Removal of Chromium from contaminated effluent and simultaneously Green Nanoparticle synthesis using Bacillus subtilis. *Malaya Journal of Biosciences*, *2014*(1), 13–18. www.malayabiosciences.com

Arunakumara, K. K. I. U., Zhang, X., & Song, X. (2008). Bioaccumulation of Pb2+ and its effects on growth, morphology and pigment contents of Spirulina(Arthrospira) platensis. *Journal of Ocean University of China*, *7*(4), 397–403. https://doi.org/10.1007/s11802-008-0397-2

Bañuelos, G. S., Vickerman, D. B., Trumble, J. T., Shannon, M. C., Davis, C. D., Finley, J. W., & Mayland, H. F. (2002). Biotransfer possibilities of selenium from plants used in phytoremediation. *International Journal of Phytoremediation*, *4*(4), 315–329. https://doi.org/10.1080/15226510208500090

Bazrafshan, E., Zarei, A. A., & Mostafapour, F. K. (2016). Biosorption of cadmium from aqueous solutions by Trichoderma fungus: kinetic, thermodynamic, and equilibrium study. *Desalination and Water Treatment*, *57*(31), 14598–14608. https://doi.org/10.1080/19443994.2015.1065764

Bedi, A., Singh, B. R., Deshmukh, S. K., Adholeya, A., & Barrow, C. J. (2018). An Aspergillus aculateus strain was capable of producing agriculturally useful nanoparticles via bioremediation of iron ore tailings. *Journal of Environmental Management*, *215*, 100–107. https://doi.org/10.1016/j.jenvman.2018.03.049

Brow, N., Rouch, D., & Lee, A. O. (1992). Copper Resistance Determinants in Bacteria. In *PLASMID* (Vol. 27).

Chakraborty, J., Mallick, S., Raj, R., & Das, S. (2018). Functionalization of Extracellular Polymers of Pseudomonas aeruginosa N6P6 for Synthesis of CdS Nanoparticles and Cadmium Bioadsorption. *Journal of Polymers and the Environment*, *26*(7), 3097–3108. https://doi.org/10.1007/s10924-018-1195-6

Chandrakant Vaigankar, D., Kumar Dubey, S., Yusuf Mujawar, S., & Kumar Mohanty, A. (2020). *Halomonas Venusta Mediated Detoxication and Biotransformation of Selenite Into Selenium Nanoparticles Exhibiting Various Biomedical Applications*. https://doi.org/10.21203/rs.3.rs-45336/v1

Chibuike, G. U., & Obiora, S. C. (2014). Heavy metal polluted soils: Effect on plants and bioremediation methods. In *Applied and Environmental Soil Science* (Vol. 2014). Hindawi Publishing Corporation. https://doi.org/10.1155/2014/752708

Cui, D., Tan, C., Deng, H., Gu, X., Pi, S., Chen, T., Zhou, L., & Li, A. (2020). Biosorption Mechanism of Aqueous Pb2+, Cd2+, and Ni2+Ions on Extracellular Polymeric Substances (EPS). *Archaea*, *2020*. https://doi.org/10.1155/2020/8891543

Dinh, N., van der Ent, A., Mulligan, D. R., & Nguyen, A. V. (2004). *Zinc and lead accumulation characteristics and in vivo distribution of Zn 2+ in the hyperaccumulator Noccaea caerulescens elucidated with fluorescent probes and laser confocal microscopy*.

Doty, S. L., James, C. A., Moore, A. L., Vajzovic, A., Singleton, G. L., Ma, C., Khan, Z., Xin, G., Jun, W. K., Jin, Y. P., Meilan, R., Strauss, S. H., Wilkerson, J., Farin, F., & Strand, S. E. (2007). Enhanced phytoremediation of volatile environmental pollutants with transgenic trees. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(43), 16816–16821. https://doi.org/10.1073/pnas.0703276104

Dursun, A. Y., Uslu, G., Cuci, Y., & Aksu, Z. (2003). *Bioaccumulation of copper(II), lead(II) and chromium(VI) by growing Aspergillus niger*. www.elsevier.com/locate/procbio

Duruibe, J. O., Ogwuegbu, M. O. C., & Egwurugwu. (2007). Heavy metal pollution and human biotoxic effects. In *International Journal of Physical Sciences* (Vol. 2, Issue 5). http://www.academicjournals.org/IJPS

Ehsan, S., Ali, S., Noureen, S., Mahmood, K., Farid, M., Ishaque, W., Shakoor, M. B., & Rizwan, M. (2014). Citric acid assisted phytoremediation of cadmium by Brassica napus L. *Ecotoxicology and Environmental Safety*, *106*, 164–172. https://doi.org/10.1016/j.ecoenv.2014.03.007

El-Naggar, N. E. A., Hamouda, R. A., Mousa, I. E., Abdel-Hamid, M. S., & Rabei, N. H. (2018). Biosorption optimization, characterization, immobilization and application of Gelidium amansii biomass for complete Pb2+ removal from aqueous solutions. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-31660-7

Eltarahony, M., Zaki, S., & Kheiralla, Z. H. (2016). *BIOCONVERSION OF SILVER AND NICKEL IONS INTO NANOPARTICLES BY ACHROMOBACTER SP. STRAIN MMT AND THEIR APPLICATION IN WASTEWATER DISINFECTION Production, purification and characterization of polysaccharide lytic enzymes of a marine isolate, Bacillus cereus NRC-20 and their application in biofilm remova View project Bio-reporter to monitor general toxicity of liquids and soils View project*. https://www.researchgate.net/publication/296585250

Fariq, A., Khan, T., & Yasmin, A. (2017). Microbial synthesis of nanoparticles and their potential applications in biomedicine. In *Journal of Applied Biomedicine* (Vol. 15, Issue 4, pp. 241–248). University of South Bohemia. https://doi.org/10.1016/j.jab.2017.03.004

Gomaa, E. Z. (2022). Microbial Mediated Synthesis of Zinc Oxide Nanoparticles, Characterization and Multifaceted Applications. *Journal of Inorganic and Organometallic Polymers and Materials*, *32*(11), 4114–4132. https://doi.org/10.1007/s10904-022-02406-w

Guo, H., Luo, S., Chen, L., Xiao, X., Xi, Q., Wei, W., Zeng, G., Liu, C., Wan, Y., Chen, J., & He, Y. (2010). Bioremediation of heavy metals by growing hyperaccumulaor endophytic bacterium Bacillus sp. L14. *Bioresource Technology*, *101*(22), 8599–8605. https://doi.org/10.1016/j.biortech.2010.06.085

Guria, M. K., Guha, A. K., & Bhattacharyya, M. (2014). A green chemical approach for biotransformation of Cr(VI) to Cr(III), utilizing Fusarium sp. MMT1 and consequent structural alteration of cell morphology. *Journal of Environmental Chemical Engineering*, *2*(1), 424–433. https://doi.org/10.1016/j.jece.2014.01.016

Hlihor, R. M., Figueiredo, H., Tavares, T., & Gavrilescu, M. (2017). Biosorption potential of dead and living Arthrobacter viscosus biomass in the removal of Cr(VI): Batch and column studies. *Process Safety and Environmental Protection*, *108*, 44–56. https://doi.org/10.1016/j.psep.2016.06.016

Hosseini, F., Hadian, M., Lashani, E., & Moghimi, H. (2023). Simultaneous bioreduction of tellurite and selenite by Yarrowia lipolytica, Trichosporon cutaneum, and their co-culture along with characterization of biosynthesized Te–Se nanoparticles. *Microbial Cell Factories*, *22*(1). https://doi.org/10.1186/s12934-023-02204-0

Hosseini, F., Lashani, E., & Moghimi, H. (2023). Simultaneous bioremediation of phenol and tellurite by Lysinibacillus sp. EBL303 and characterization of biosynthesized Te nanoparticles. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-28468-5

Ilyinskaya, E., Mason, E., Wieser, P. E., Holland, L., Liu, E. J., Mather, T. A., Edmonds, M., Whitty, R. C. W., Elias, T., Nadeau, P. A., Schneider, D., McQuaid, J. B., Allen, S. E., Harvey, J., Oppenheimer, C., Kern, C., & Damby, D. (2021). Rapid metal pollutant deposition from the volcanic plume of Kīlauea,

Hawai'i. *Communications Earth and Environment*, *2*(1). https://doi.org/10.1038/s43247-021-00146-2

Jadoun, S., Chauhan, N. P. S., Zarrintaj, P., Barani, M., Varma, R. S., Chinnam, S., & Rahdar, A. (2022). Synthesis of nanoparticles using microorganisms and their applications: a review. In *Environmental Chemistry Letters* (Vol. 20, Issue 5, pp. 3153–3197). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s10311-022-01444-7

Jalilvand, N., Akhgar, A., Alikhani, H. A., Rahmani, H. A., & Rejali, F. (2020). Removal of Heavy Metals Zinc, Lead, and Cadmium by Biomineralization of Urease-Producing Bacteria Isolated from Iranian Mine Calcareous Soils. *Journal of Soil Science and Plant Nutrition*, *20*(1), 206–219. https://doi.org/10.1007/s42729-019-00121-z

Javaid, A., Bajwa, R., Shafique, U., & Anwar, J. (2011). Removal of heavy metals by adsorption on Pleurotus ostreatus. *Biomass and Bioenergy*, *35*(5), 1675–1682. https://doi.org/10.1016/j.biombioe.2010.12.035

Kavita, B., & Keharia, H. (2012). Biosorption potential of Trichoderma gamsii biomass for removal of Cr(VI) from electroplating industrial effluent. *International Journal of Chemical Engineering*. https://doi.org/10.1155/2012/305462

Kiran, I., Akar, T., & Tunali, S. (2005). Biosorption of Pb(II) and Cu(II) from aqueous solutions by pretreated biomass of Neurospora crassa. *Process Biochemistry*, *40*(11), 3550–3558. https://doi.org/10.1016/j.procbio.2005.03.051

Kour, D., Kaur, T., Devi, R., Yadav, A., Singh, M., Joshi, D., Singh, J., Deep, &, Suyal, C., Kumar, A., Rajput, V. D., Ajar, &, Yadav, N., Singh, K., Singh, J., Sayyed, R. Z., Naveen, &, Arora, K., Anil, &, & Saxena, K. (2021). Beneficial microbiomes for bioremediation of diverse contaminated environments for environmental sustainability: present status and future challenges. *Environmental Science Pollution Research International*. https://doi.org/10.1007/s11356-021-13252-7/Published

Kumar Sharma, R., & Agrawal, M. (2005). Biological effects of heavy metals: An overview. *Journal of Environmental Bioloy*. http://www.geocities.com/Lenviron_biol/J.Environ.BioI.26

Kurniati, E., Arfarita, N., Imai, T., Higuchi, T., Kanno, A., Yamamoto, K., & Sekine, M. (2014). Potential bioremediation of mercury-contaminated substrate using filamentous fungi isolated from forest soil. *Journal of Environmental Sciences (China)*, *26*(6), 1223–1231. https://doi.org/10.1016/S1001-0742(13)60592-6

L, F.-T. J. (2010). *Environmental geochemistry of ancient volcanic ashes*.

Liaquat, F., Munis, M. F. H., Haroon, U., Arif, S., Saqib, S., Zaman, W., Khan, A. R., Shi, J., Che, S., & Liu, Q. (2020). Evaluation of metal tolerance of fungal strains isolated from contaminated mining soil of Nanjing, China. *Biology*, *9*(12), 1–12. https://doi.org/10.3390/biology9120469

Liu, Z., Chen, M., Lin, M., Chen, Q., Lu, Q., Yao, J., & He, X. (2022). Cadmium Uptake and Growth Responses of Seven Urban Flowering Plants: Hyperaccumulator or Bioindicator? *Sustainability (Switzerland)*, *14*(2). https://doi.org/10.3390/su14020619

Ma, Q., Han, L., Zhang, J., Zhang, Y., Lang, Q., Li, F., Han, A., Bao, Y., Li, K., & Alu, S. (2019). Environmental risk assessment of metals in the volcanic soil of Changbai mountain. *International Journal of Environmental Research and Public Health*, *16*(11). https://doi.org/10.3390/ijerph16112047

Mane, P. C., & Bhosle. (2012). Bioremoval of Some Metals by Living Algae Spirogyra sp. and Spirullina sp. from aqueous solution. *Int. J. Environ. Res*, *6*(2), 571–576. www.SID.ir

McGrath, S. P., & Zhao, F. J. (2003). Phytoextraction of metals and metalloids from contaminated soils. In *Current Opinion in Biotechnology* (Vol. 14, Issue 3, pp. 277–282). Elsevier Ltd. https://doi.org/10.1016/S0958-1669(03)00060-0

Mishra, A., & Malik, A. (2012). Simultaneous bioaccumulation of multiple metals from electroplating effluent using Aspergillus lentulus. *Water Research*, *46*(16), 4991–4998. https://doi.org/10.1016/j.watres.2012.06.035

Mohapatra, R. K., Parhi, P. K., Pandey, S., Bindhani, B. K., Thatoi, H., & Panda, C. R. (2019). Active and passive biosorption of Pb(II)using live and dead biomass of marine bacterium Bacillus xiamenensis PbRPSD202: Kinetics and isotherm studies. *Journal of Environmental Management*, *247*, 121–134. https://doi.org/10.1016/j.jenvman.2019.06.073

Murali, M., Gowtham, H. G., Shilpa, N., Singh, S. B., Aiyaz, M., Sayyed, R. Z., Shivamallu, C., Achar, R. R., Silina, E., Stupin, V., Manturova, N., Shati, A. A., Alfaifi, M. Y., Elbehairi, S. E. I., & Kollur, S. P. (2023). Zinc oxide nanoparticles prepared through microbial mediated synthesis for therapeutic applications: a possible alternative for plants. *Frontiers in Microbiology*, *14*. https://doi.org/10.3389/fmicb.2023.1227951

Musa Jibrin, A., Adebayo Oyewole, O., Gauis Yakubu, J., Hussaini, A., & Chidi Egwim, E. (2020). Heavy metals biosorption by urease producing Lysinibacillus fusiformis 5B. *European Journal of Biological Research Research Article European Journal of Biological Research*, *10*(4), 326–335. https://doi.org/10.5281/zenodo.4023161

Mwandira, W., Nakashima, K., Kawasaki, S., Arabelo, A., Banda, K., Nyambe, I., Chirwa, M., Ito, M., Sato, T., Igarashi, T., Nakata, H., Nakayama, S., & Ishizuka, M. (2020). Biosorption of Pb (II) and Zn (II) from aqueous solution by Oceanobacillus profundus isolated from an abandoned mine. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-78187-4

Narolkar, S., Jain, N., & Mishra, A. (2022). Biosorption of Chromium by Fungal Strains Isolated from Industrial Effluent Contaminated Area. *Pollution*, *8*(1), 159–168. https://doi.org/10.22059/POLL.2021.326818.1132

Nourbakhsh, M., Sag, " Y, &er, " D, Aksu, Z., Kutsal, T., & Ca@ar, & A. (1994). A Comparative Study of Various Biosorbents for Removal of Chromium(V1) Ions from Industrial Waste Waters. In *Process Bioc&rni~* (Vol. 29).

Pandey, S., Awasthee, N., Shekher, A., Rai, L. C., Gupta, S. C., & Dubey, S. K. (2021). Biogenic synthesis and characterization of selenium nanoparticles and their applications with special reference to antibacterial, antioxidant, anticancer and photocatalytic activity. *Bioprocess and Biosystems Engineering*, *44*(12), 2679–2696. https://doi.org/10.1007/s00449-021-02637-0

Pavan Kumar Gautam, & Ravindra Kuma, S. B. M. C. C. J. D. P. (2022). *Heavy metals : sources, toxicity and remediation techniques*.

Pham, T. L. (2019). Removal and bioaccumulation of copper by the freshwater green alga Scenedesmus sp. *Vietnam Journal of Science, Technology and Engineering*, *61*(2), 65–70. https://doi.org/10.31276/VJSTE.61(2).65-70

Pham, T. L., Dao, T. S., Bui, H. N., Pham, T. K. N., Ngo, T. T. H., & Bui, H. M. (2020). Lipid production combined with removal and bioaccumulation of Pb by Scenedesmus sp. Green Alga. *Polish Journal of Environmental Studies*, *29*(2), 1785–1791. https://doi.org/10.15244/pjoes/109277

Puyen, Z. M., Villagrasa, E., Maldonado, J., Diestra, E., Esteve, I., & Solé, A. (2012). Biosorption of lead and copper by heavy-metal tolerant Micrococcus luteus DE2008. *Bioresource Technology*, *126*, 233–237. https://doi.org/10.1016/j.biortech.2012.09.036

Ramanathan, R., Field, M. R., O'Mullane, A. P., Smooker, P. M., Bhargava, S. K., & Bansal, V. (2013). Aqueous phase synthesis of copper nanoparticles: A link between heavy metal resistance and nanoparticle synthesis ability in bacterial systems. *Nanoscale*, *5*(6), 2300–2306. https://doi.org/10.1039/c2nr32887a

Ranieri, E., Fratino, U., Petrella, A., Torretta, V., & Rada, E. C. (2016). Ailanthus Altissima and Phragmites Australis for chromium removal from a contaminated soil. *Environmental Science and Pollution Research*, *23*(16), 15983–15989. https://doi.org/10.1007/s11356-016-6804-0

Ranjani, K. A. (2019). *BIOSORPTION OF HEAVY METAL CHROMIUM BY LIVE AND DEAD FUNGIISOLATED FROM LEATHER TANNERY*.

Rinanti, A., Fachrul, M. F., Hadisoebroto, R., & Silalahi, M. D. S. (2018). Biosorption of Cu (Ii) by Scenedesmus obliquus: Optimization in Phovasoli Haemotococcus Medium. *International Journal of GEOMATE*, *15*(52), 45–52. https://doi.org/10.21660/2018.52.24521

Robinson, B. H., Lombi, E., Zhao, F. J., & McGrath, S. P. (2003). Uptake and distribution of nickel and other metals in the hyperaccumulator Berkheya coddii. *New Phytologist*, *158*(2), 279–285. https://doi.org/10.1046/j.1469-8137.2003.00743.x

Rouchalova, D., Rouchalova, K., Janakova, I., Cablik, V., & Janstova, S. (2020). Bioleaching of iron, copper, lead, and zinc from the sludge mining sediment at different particle sizes, ph, and pulp density using acidithiobacillus ferrooxidans. *Minerals*, *10*(11), 1–28. https://doi.org/10.3390/min10111013

Rugnini L, Costa G, Congestri R, Antonaroli S, & Bruno L. (2018). *PHOSPHORUS AND METAL REMOVAL COMBINED WITH LIPID PRODUCTION BY THE GREEN MICROALGA Desmodesmus sp.: AN INTEGRATED APPROACH*.

Salvadori, M. R., Ando, R. A., Oller Do Nascimento, C. A., & Corrêa, B. (2014). Intracellular biosynthesis and removal of copper nanoparticles by dead biomass of yeast isolated from the wastewater of a mine in the Brazilian Amazonia. *PLoS ONE*, *9*(1). https://doi.org/10.1371/journal.pone.0087968

Sankarammal, M., Thatheyus, A., & Ramya, D. (2014). Bioremoval Of Cadmium Using Pseudomonas fluorescens. *Open Journal of Water Pollution and Treatment*, *2014*(2), 92–100. https://doi.org/10.15764/WPT.2014.02010

Sathyavathi, S., Manjula, A., Rajendhran, J., & Gunasekaran, P. (2013). Biosynthesis and characterization of mercury sulphide nanoparticles produced by Bacillus cereus MRS-1. In *Indian Journal of Experimental Biology* (Vol. 51).

Shamshad, I., Khan, S., Waqas, M., Asma, M., Nawab, J., Gul, N., Raiz, A., & Li, G. (2016). Heavy metal uptake capacity of fresh water algae (Oedogonium westti) from aqueous solution: A mesocosm research. *International Journal of Phytoremediation*, *18*(4), 393–398. https://doi.org/10.1080/15226514.2015.1109594

Sinha, A., Pant, K. K., & Khare, S. K. (2012). Studies on mercury bioremediation by alginate immobilized mercury tolerant Bacillus cereus cells. *International Biodeterioration and Biodegradation*, *71*, 1–8. https://doi.org/10.1016/j.ibiod.2011.12.014

Sobhy, E. E., Abd, E.-R. R. E.-S., & Perihan, S. A. (2014). Accumulation of some heavy metals by metal resistant avirulent Bacillus anthracis PS2010 isolated from Egypt. *African Journal of Microbiology Research*, *8*(12), 1266–1276. https://doi.org/10.5897/ajmr2013.6551

Stefanowicz, A. M., Kapusta, P., Zubek, S., Stanek, M., & Woch, M. W. (2020). Soil organic matter prevails over heavy metal pollution and vegetation as a factor shaping soil microbial communities at historical Zn–Pb mining sites. *Chemosphere*, *240*. https://doi.org/10.1016/j.chemosphere.2019.124922

Sundararaju, S., Arumugam, M., & Bhuyar, P. (2020). Microbacterium sp. MRS-1, a potential bacterium for cobalt reduction and synthesis of less/non-toxic cobalt oxide nanoparticles (Co3O4). *Beni-Suef University Journal of Basic and Applied Sciences*, *9*(1). https://doi.org/10.1186/s43088-020-00070-y

Tyagi, S., Tyagi, P. K., Gola, D., Chauhan, N., & Bharti, R. K. (2019). Extracellular synthesis of silver nanoparticles using entomopathogenic fungus: characterization and antibacterial potential. *SN Applied Sciences*, *1*(12). https://doi.org/10.1007/s42452-019-1593-y

Uddin, A., & Lall, A. M. (2019). Phycoremediation of Heavy Metals by Botryococusbrurauni. *Biosciences, Biotechnology Research Asia*, *16*(1), 129–133. https://doi.org/10.13005/bbra/2730

Upadhyay, A. K., Singh, L., Singh, R., Singh, D. P., & Saxena, G. (2022). Bioaccumulation and Toxicity of As in the Alga Chlorococcum sp.: Prospects for As Bioremediation. *Bulletin of Environmental Contamination and Toxicology*, *108*(3), 500–506. https://doi.org/10.1007/s00128-020-02964-0

Urík, M., Čerňanský, S., Littera, P., & Kolencik, M. (2010). *Biosorption and Bioaccumulation of Thallium(I) and Its Effect on Growth of Neosartorya fischeri Strain*. https://www.researchgate.net/publication/222106450

Van Aken, B. (2009). Transgenic plants for enhanced phytoremediation of toxic explosives. In *Current Opinion in Biotechnology* (Vol. 20, Issue 2, pp. 231–236). https://doi.org/10.1016/j.copbio.2009.01.011

Verma, J., Bhatt, A., & Kumar Agrawal, P. (2016). International Peer-Reviewed Journal ISSN 2321 3655 Oct. In *Jour. Env. Res* (Vol. 4, Issue 2). http://www.sciencebeingjournal.com

Vidali, M. (2001a). Bioremediation. An overview*. In *Pure Appl. Chem* (Vol. 73, Issue 7).

Vidali, M. (2001b). Bioremediation. An overview*. In *Pure Appl. Chem* (Vol. 73, Issue 7).

Vigara Fernández, J., Author, F., Leon-Vaz, A., María Vega, J., Carfagna, S., Andersen-Ranberg, J., & Mostafa, I. (2021). *Aquatic Toxicology Impact of heavy metals in the microalga Chlorella sorokiniana and assessment of its potential use in cadmium bioremediation Expert in microalgae and metallic stress Expert in microalgae and plants Expert in microalgae Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation*.

Wen, J., Wu, Y., Li, X., Lu, Q., Luo, Y., Duan, Z., & Li, C. (2021). Migration characteristics of heavy metals in the weathering process of exposed argillaceous sandstone in a mercury-thallium mining area. *Ecotoxicology and Environmental Safety*, *208*. https://doi.org/10.1016/j.ecoenv.2020.111751

White, C., Wilkinson, S. C., & Gadd, G. M. (1995). The Role of Microorganisms in Biosorption of Toxic Metals and Radionuclides. In *International Biodeterioration & Biodegradation*.

Yadav, M., & Sharma, | Poonam. (2023). *Current Eco-friendly and Sustainable Methods for Heavy Metals Remediation of Contaminated Soil and Water: Special Emphasis on Use of Genetic Engineering and Nanotechnology*. https://doi.org/10.22059/poll.2023.345177.1522

Ye, J., Xiao, H., Xiao, B., Xu, W., Gao, L., & Lin, G. (2015). Bioremediation of heavy metal contaminated aqueous solution by using red algae Porphyra leucosticta. *Water Science and Technology*, *72*(9), 1662–1666. https://doi.org/10.2166/wst.2015.386

Zhang, T., & Zhang, H. (2022). Microbial Consortia Are Needed to Degrade Soil Pollutants. In *Microorganisms* (Vol. 10, Issue 2). MDPI. https://doi.org/10.3390/microorganisms10020261

Zouboulis, A. I., Loukidou, M. X., & Matis, K. A. (2004). Biosorption of toxic metals from aqueous solutions by bacteria strains isolated from metal-polluted soils. *Process Biochemistry*, *39*(8), 909–916. https://doi.org/10.1016/S0032-9592(03)00200-0

Žvab, U., Kukulin, D. S., Fanetti, M., & Valant, M. (2021). Bioremediation of copper polluted wastewater-like nutrient media and simultaneous synthesis of stable copper nanoparticles by a viable green alga. *Journal of Water Process Engineering*, *42*. https://doi.org/10.1016/j.jwpe.2021.102123

IOP Conference Series: Earth and Environmental Science

PAPER • OPEN ACCESS

Buildings of Distinct Plan Shapes Under Wind Load: Wind Interference Effects

To cite this article: Supriya Pal *et al* 2024 *IOP Conf. Ser.: Earth Environ. Sci.* **1326** 012002

View the article online for updates and enhancements.

You may also like

- Can CGCMs under CMIP5/6 simulate present-day sea level rise in western Maritime Continent?
Y S Djamil, A Maharani, T Solihuddin et al.
- Exploring the Potential of High-Fatty Acid Content Oils for Biodiesel Production: A Catalyst-Free Approach
MD Solikhah, SS Wirawan, AT Sugiarto et al.
- Dynamic Modelling, Simulation, and Sensitive Analysis of Lead Removal in a Fixed-Bed Adsorption Column using Waste-Based Materials
Mohammad Gheibi, Stanisaw Wacawek, Choe Peng Leo et al.

This content was downloaded from IP address 14.139.251.98 on 01/07/2024 at 05:35

Buildings of Distinct Plan Shapes Under Wind Load: Wind Interference Effects

Supriya Pal*[1], Ritu Raj[2], Alfia Bano[3]
[1] Faculty Civil Engineering Department, National Institute of Technology, Raipur, India.
[2]Faculty of Civil Engineering Department, Delhi Technological University, Delhi, India.
[3]Faculty of Civil Engineering Department, National Institute of Technology, Raipur, India.

Corresponding Author: supriya8788@gmail.com

Abstract. This paper presents the results of the experimental investigations carried out on two intricate plan-shaped buildings to evaluate the aerodynamic performance of the building. The intention is to explore the associated environment resulting from wind forces in both solitary and obstructing situations using a prototype model at a scale of 1:300 with a prototype having 180m height. Experimental tests were carried out at the boundary layer wind tunnel considering isolated conditions with 0° to 180° isolated wind directions and four interference conditions of twin buildings (separately for both the models). The research has been undertaken in Terrain Category II, which is the most widely available terrain in India. A comparative assessment of the two models is performed in an attempt to comprehend the various consequences of diverse conditions that may emerge in real-life situations, as well as the discrepancies among them. Experimental results of wind pressure coefficients of Model-1 and Model-2 show good agreement with various wind incidence conditions with the minute difference in the magnitudes of Cp. On the basis of wind tunnel studies, it is distinguished that the performance of Model-2 is better than Model-1in both isolated as well as interference conditions for all wind incidences and orientations respectively.

Keywords: Tall buildings; wind directions; interference factor; mean pressure-coefficients

1. Introduction

Impelled by the developing need of the present era, the trend calls for the construction of tall buildings. Buildings, according to Hassanli et al. (2017), do provide shelter and also operate as a barrier between both the outdoor and inside surroundings, as well as the residents who live within. Amin and Ahuja (2010) said the barrier i.e., the tall structures are generally slender in nature and are induced by wind load and can cause discomfort to the inhabitants on a psychological level and the symptoms may range from concern, anxiety, and fear to headaches. Kalehsar and Khodaie (2017) and (2018) modeled a structural design for super-tall buildings for the reduction of excessive vibration due to wind loads by inducing damping devices. An ample amount of information is available with different international standards (AS/ NZS: 1170.2:2011, ASCE: 7-02 (2002), IS: 875 (Part 3): 2015) for regular symmetrical cross- sectional plan shape buildings with similar cross-section throughout its height around the world. As necessary as the study of aerodynamic effects on the isolated structures is, seldom these structures are constructed as a unit. Only a small amount of investigation has been executed out over indifferent planned structures. Chakraborty et al. (2013), Cheng et al. (2015), Li (2017), Gomes et al. (2005), Bhattacharyya et al. (2014) and Paul and Dalui (2016) studied through experiments the wind pressures on '+' shaped, "H" shaped "L" shaped, L‟ and „U‟ plan shaped, L- and T-shaped, E- plan shaped and Z-plan shaped models of tall structures respectively, for wind angles varying wind incidence angles. The

Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

results exhibit a huge difference in wind-induced pressures on windward as well as leeward faces as compared to a regular cubical model. Although the experimental study gives a fair picture of the field conditions and the results, some field studies have also been carried out for producing extreme conditions. Hangan et al, (2017), had analyzed new full-scale measurement techniques such as Doppler Radars and LiDARs to allow a better portrayal of events such as tornadoes and downbursts. Chuang and Spence (2017) had done a full-scale case study of a 40-story four-span plane steel frame and including parameters like wind hazard analysis, aerodynamic analysis, structural analysis, damage analysis, non-collapse loss assessment, collapse assessment, dynamic shakedown and collapse of wind excited system and collapse loss assessment.

Solari (2017) studied the topography effects, atmospheric turbulence and aerodynamic wind actions; found that the dynamic wind in nature as stated by Hu et al. (2017), causes huge damages that frequently may lead to collapse of the buildings. In order to reduce the induced wind loads, dynamic responses and the adverse effects, we need a broad study over other parameters such as size, shape, edge and or corner modifications of a building. The formation of wind loads on high-rise structures is known to be significantly influenced by the external shapes of tall buildings along their height. The optimized shape of a building can be chosen by altering the flow pattern around a building by modifying the building shape (Ahmed et al. (2017) and Alminhana (2018)). Zheng et al. (2018) have experimentally investigated the aerodynamic performance of tall building models with modified corners (chamfered and recessed). Stathopoulos and Zhou (1993) pictured the wind pressures on surfaces of various buildings with sharp corners by adopting numerical simulation methods. Through the studies it has been verified that the recessed type corners are more efficient in the reduction of the lift forces and also, chamfered type corners prove to be more efficient in the reduction of the drag force. Kim and Kanda (2013) investigated the spatial-temporal characteristics of pressure fluctuations on tapered and set-back tall buildings. Silva et al, (2018) studied conventional roof and roof edge shapes for building roof shape optimization for the wind energy exploitation for both turbulence intensity and wind velocity.

Most of the buildings are constructed in clusters with different arrangements among themselves as well as with the upstream wind. Due to the close proximity of another building, the wind pressure coefficients of such structures vary from that of the isolated standing structure. The interference effect between two or more buildings in close proximity is important in the field of building aerodynamics (Khanduri et al. 1997). Wind pressure on the main structure may grow or fall as a result of the adjacent buildings' obstruction, depending on a number of factors (Chitra et al. 2017). Tse et al, (2017) have experimentally investigated the surrounding wind conditions of building arrays in twisted wind flows, Xu et al (2017) has performed a comparative study between 400m and 200m tall buildings experimentally and investigated the pedestrian level winds around 40 different configurations super tall building models through wind tunnel tests, Amin and Ahuja (2012) has calculated the mean interference effects between two rectangular buildings (‘L’and ‘T’ plan shape) located in close proximity by wind tunnel test and Ming (2010) carried out wind tunnel tests on 27 typical tall building models and analysed the characteristics of wind-induced pressures and forces on these buildings. Through these studies, as concluding remarks, it was said that the building height and passage width are the most influential parameters in determining acceptable wind conditions near building arrays. Many other analytical, numerical, and or experimental studies has been carried out to evaluate wind loads on buildings in interference conditions for various indifferent structure with multiple arrangements among themselves (Kheyari and Dalui (2015), Kar and Dalui (2016), Bairagi and Dalui (2014), Mukherjee and Bairagi (2018), Gurjar and Amin (2017)). Li et al. (2017) concluded that there is no analytical formula available for evaluation of the wind effects on irregularly shaped tall buildings with or without interfering with buildings in close vicinity.

Although, a number of studies have been done to check the induced wind pressure on the faces of indifferent plan shape buildings, to date, no comparative study between two complex plan-shaped structures has been indexed. In this paper, a number of parameters (such as complex shape and change in shape) are taken into account for the experimental study through BLWT. A thorough

study of two complex structures, Model-1 (M-1)-Figure. 2(a) and Model-2 (M-2) Figure. 2(b) has been carried out. The results obtained from the pressure distribution on Face-A only in isolation and interference have been explained in detail due to the limited space of the paper. Further, a comparative study has been done for a better understanding of the effects of minor change the in plan of building towards the associated pressure.

2. Methodology

2.1 Experimental Configuration Aspects

The experiment has been carried out in a BLWT (Boundary Layer Wind Tunnel) with a cross sectional area of 4 x 10^6 mm^2 and a total length of 15 x 10^3 mm. The fluid motion is lined by establishing a square shape honey comb at the tunnel's inlet, which also generates the required vortex flow. Terrain has been developed for the experimental study by placing the physical barriers at the upstream of the wind tunnel. With the assistance of suction from blower fan, the wind continually travels through the tunnel at a constant speed of 10 m/sec.

On each surface of the building model, the flow separation-induced pressure variation is evaluated. A pressure tapping built of 1 mm steel tubing is used to detect pressure. This resulting pressure is gauged and connected to a Baron instrument via a pressure transducer that is connected to each pressure taping. On a one-meter-long probe, the "TESTO - 480" Baratron Pressure Transducer can detect relatively modest differential head Wind speeds in the tunnel at various heights. The variation in mean wind profile velocity and turbulence intensity recorded above the turn table at the downstream end of the wind tunnel is shown in Figure 1. The mean wind speed profile inside the wind tunnel follows the Power Law with an Index, $\alpha = 0.22$ for the present study. The turbulent intensity of the flow is found 12% near the floor of the tunnel.

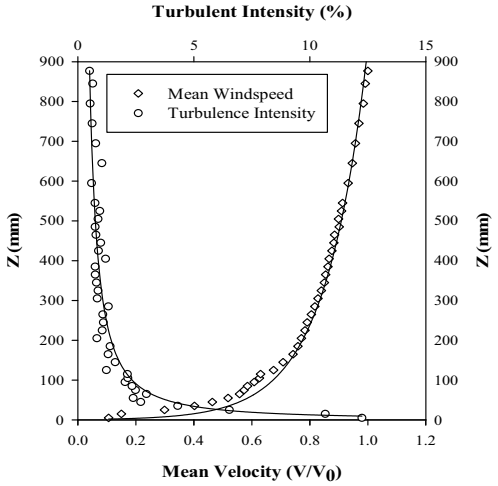


Figure 1. Mean wind speed and turbulence intensity profile

2.2 Details of models

The building models are scaled at 1:300 length scale. For the purposes of comparison, the plan area and height of the two models were maintained the same at 40,000 mm^2 and 600 mm, respectively. Figures 2(a) and 2(b) for M-1 and M-2, respectively, show all the specifications for the plan and elevation views as well as the spots of the pressure measurements of the models. Despite the fact that the area has remained the same, it is easy to see how the plan has altered.

Both the models are instrumented with similar positions of pressure taping points uniformly in order to check for the complete distribution as well for relating the results among both the models for pressure distribution at each face separately. However, in this paper the pressure values in the

form of pressure coefficients are pinpointed for Face-A only for both the models for the wind incident angles (0^0, 30^0, 60^0, 90^0, 120^0, 150^0 and 180^0) for isolated conditions, Figure. 3(a) and Figure. 3 (b) for M-1 and M-2 respectively.

Cp for four interference conditions for both the models have been experimentally calculated. For each location of pressure measurements, the pressure coefficient is established as: IS: 875 (Part 3), 2015.

$$C_p = \frac{\text{Pressure at any point } (P_a)}{0.6\ V_z^2}$$

Where, Vz is the design wind speed in m/s (10 m/s for this experiment).

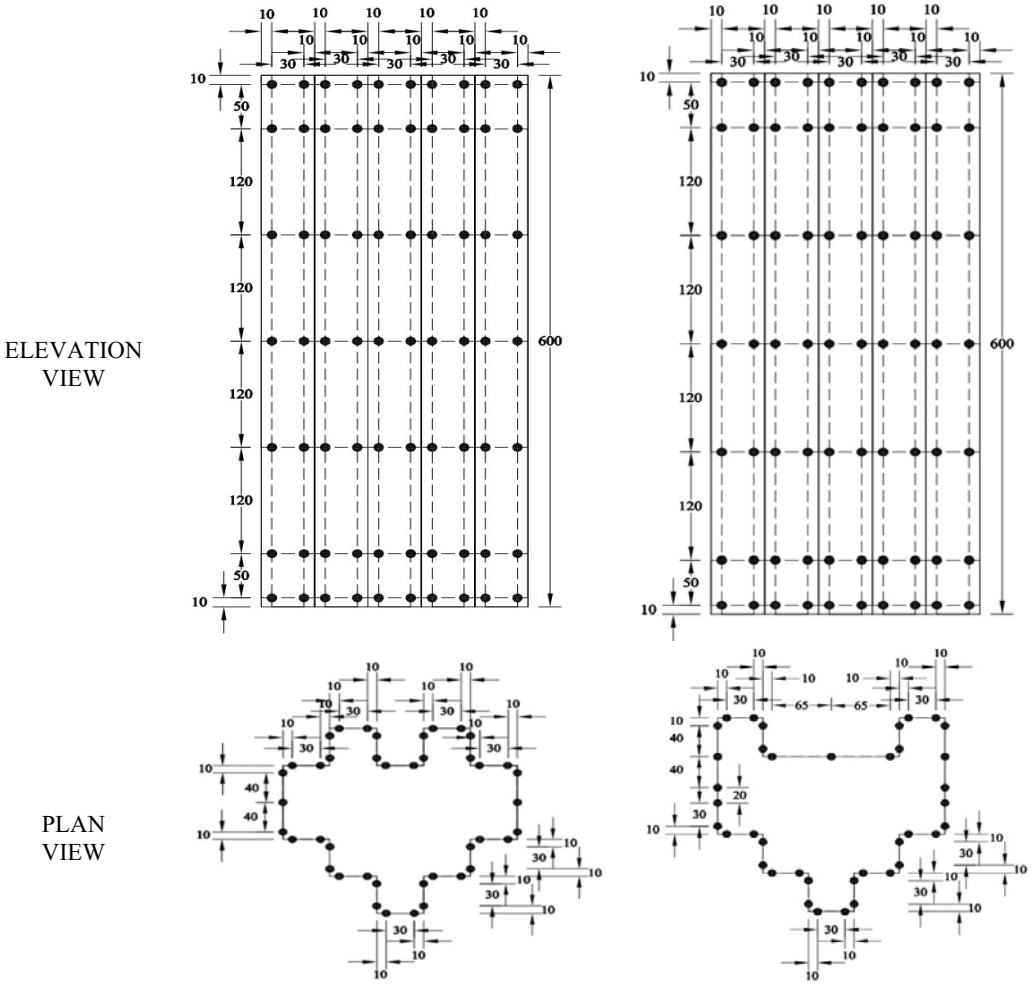
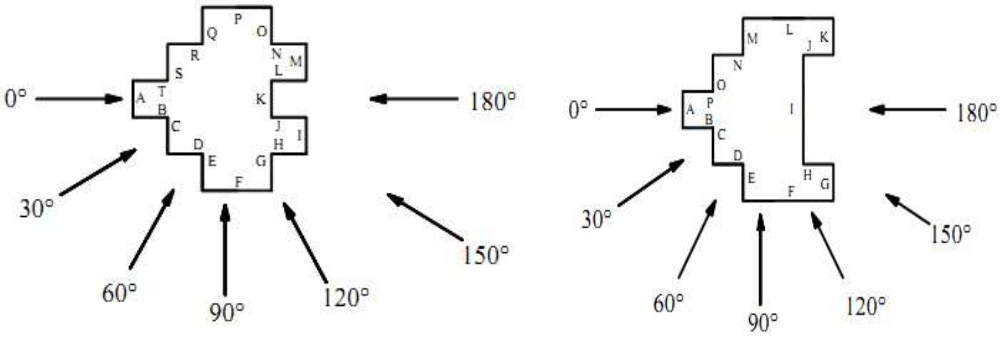


Figure 2. Pressure taping layout of building models

Both the models have more than 270 pressure taps at seven different height levels of 10mm, 60mm, 180mm, 300mm, 420mm, 540mm and 590mm from the bottom up to get a better image of the pressure distribution on all sides and faces of the models.

The objective of the study is to discover how the building's external shape's slight alterations affect the flow streamlines. For both the models of isolated and interferential situations, a thorough study of Face-A has been demonstrated. Additionally, average Cp magnitudes for each face for both models are performed, and a full discussion is conducted. Interference Factors are used to study the impact of having a building model nearby, and the results are then thoroughly described.



(a) M-1 (b) M-2

Figure. 3 Various wind incidences (Isolated Condition)

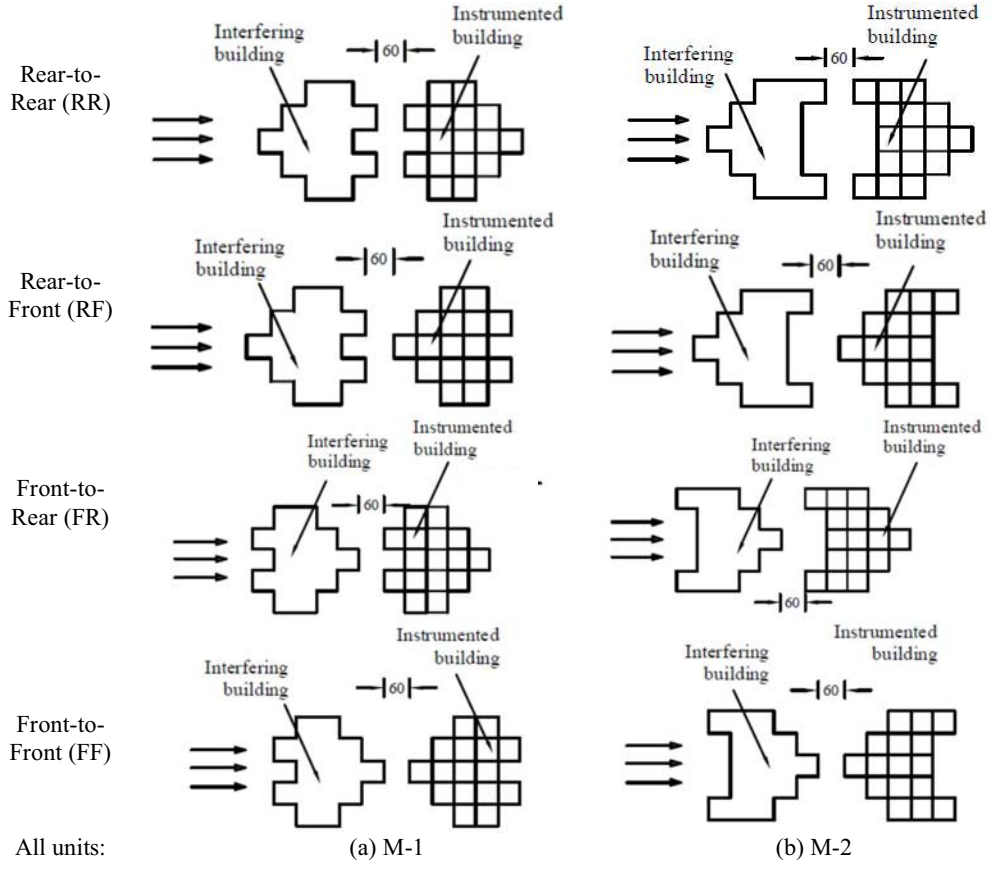


Figure. 4 Various Interference Conditions (Unit: mm)

3. Analysis Of Results: Isolated Condition

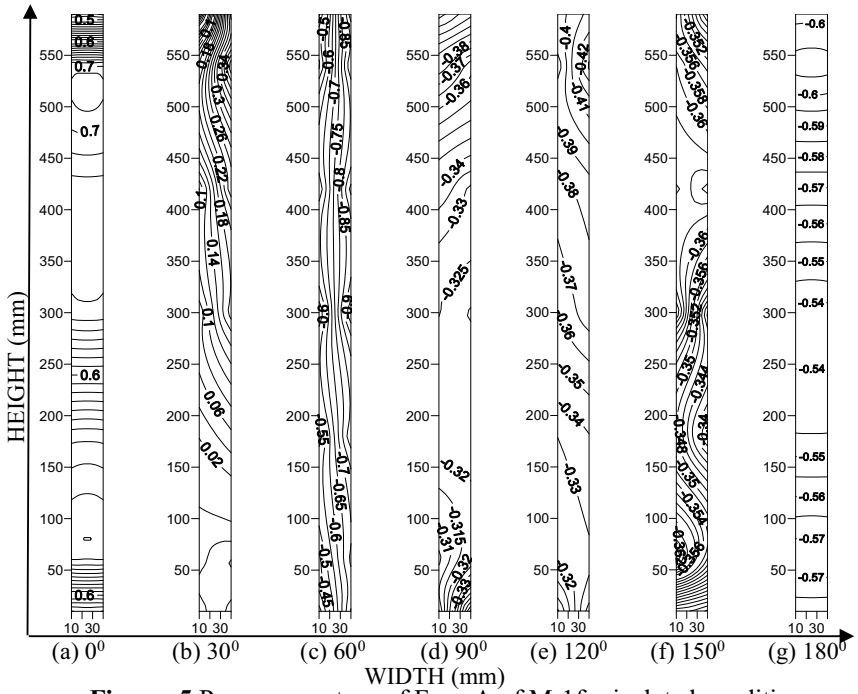


Figure. 5 Pressure contour of Face-A of M-1for isolated condition

The distribution of mean pressure coefficients is shown in Figure. 5 for isolated conditions with wind incidence angle from 0° to 180° at an interval of 30°. For 0° wind incidence the pressure is uniformly distributed from 0.51 to 0.67 on the 1/3rd to 1/2nd height of the building model due to direct undeviating wind force. At 30° wind incidence condition the wind is getting twisted along the horizontal centerline and Face-A is now not in the direction of undeviating wind force, due to this condition the magnitudes of C_p are seen to have lower magnitudes as compared to 0^0 wind incidence condition. Further, at 60° angle of incidence suction is generated throughout the face with a magnitude ranging between -0.50 to -0.85. Suction is created because of the skewed angle of wind incidence along with the interference of nearby faces thus flow separation is taking place at Face-A. At 90°, the downstream wind direction is further getting away from Face-A and disturbances can be observed to form at the upper ½ and bottom 1/3rd region only, at 1/2nd to 1/3rd distance form half height of model no pressure effect is seen. At 120° the wind is twisting but the direction of it is just opposite that of at 90° as well the fluctuation is also not too high as compared to other wind incidences. At 150° huge fluctuation of wind pressure and large twisting of the wind can be seen as the Face-A falling into the wake region. At 180° wind incidence, the distribution of Cp is observed to get uniformly distributed over the height of the model due to the symmetrical plan shape as wind is getting equally distributed at both sides after it is incident at Face-K and gathering at Face-A from both sides. The maximum C_p of 0.71 is found at 1/4th height of the model from top for 0° wind incidence and the minimum C_p of -0.31 is found at 1/6th of the height of the model from the bottom at 90° wind incidence.

Figure. 6 shows mean pressure distribution of M-2 for isolated conditions at wind incidence angles 0° to 180° at an interval of 30°. The flow as well as the plan of M-1 is symmetrical in 0° and 180°. The magnitude of C_p is varying from 0.40 to 0.69 at 0° whereas, the variation of C_p at 180° is from -0.47 to -0.59, both the magnitudes show large similarities to Cp of M-1 at 0° and 180° respectively. At 0° Face-A is at direct incident of wind flow and at 180° the face is at the leeward side and falls into wake region. The maximum and minimum Cp's also falls at 0° and 180° with, 0.68 and -0.47 magnitudes of Cp respectively. However, no influence of wind is observed at height 400mm to 520mm and 150mm to 320mm respectively for 0° and 180° wind incidences. At 30° of

wind incidence for bottom 1/6th height no effect of induced pressure is observed. Above this height the pressure is rising from 0 to 0.30 where the wind can be observed to twist at upper 1/6th height of the model with fewer fluctuations.

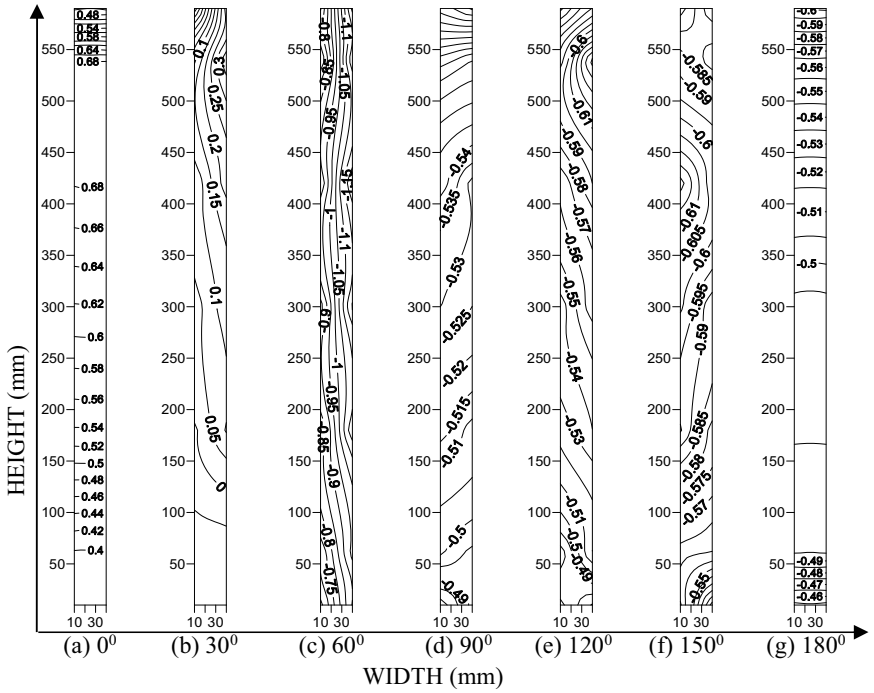
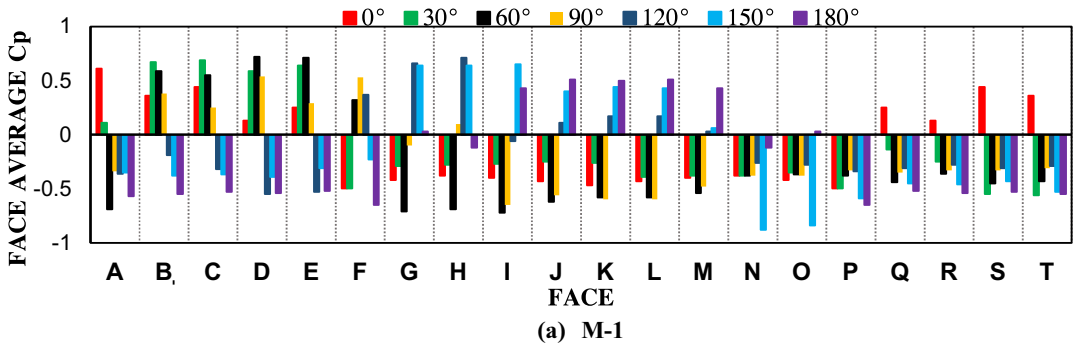
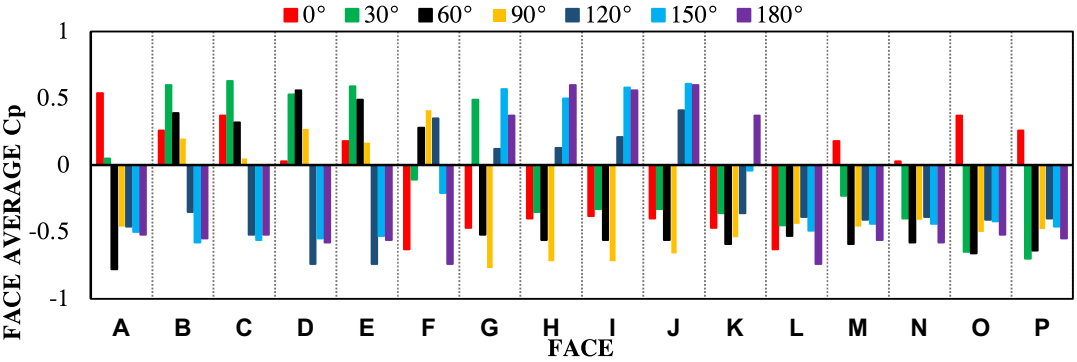


Figure. 6 Pressure contour of Face-A of M-2 for isolated condition



(a) M-1



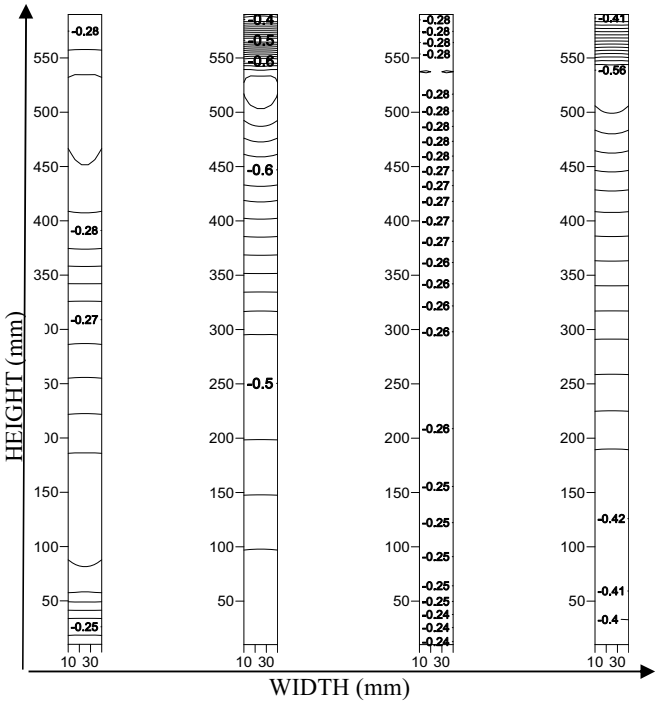
(b) **M-2**

Figure. 7 Face Average Cp distribution al the periphery (Isolated condition) (a) Model-1; (b) Model-2

At 60° wind incidence (Figure 6), large fluctuations at the Face-A along the width can be observed which is distributed throughout the height of the model. The least magnitude of suction -0.75 can be observed at bottom height of the model which is constantly rising till -1.15 magnitudes at 3/4th height of model. The Cp continues to fall until it reaches the lowest magnitude of -0.75 at the model's highest point. While suction is still produced at 90°, the change in Cp throughout the height is reduced (0.48 to 0.54), with the least magnitude at the bottom and top and the greatest magnitude of 0.54 at 420 mm. From a height of 420mm to 600mm the fluctuations are rapid which shows upwash phenomena of the wind flow. Although, the direction of twisting wind at 90° is opposite from that at 30° but the magnitude and variation at the Face-A is similar to that at 120°, with the magnitude varying from -0.49 to -0.61. At 7/8th height of the model a large fluctuation and thus rotation of wind can be seen with 0.62 at the center of the rotation.

Maximum face value of M-1(Figure. 7(a)) is about 0.71 at isolated condition is observed at 60^0 (Face-E) and 120^0 (Face-H) direction of wind incidences simultaneously both due to the depressed positioning of the faces and the wind is getting trapped at between the depressed edges at both the faces. From cladding design point of view isolated building orientations at 30^0, 60^0 and 120^0 angles of wind incidences should be given extra care while designing whereas the building orientation at 0^0 wind incidence is best suited as of generates average magnitudes of face values among all.

At M-2 (Figure. 7(b)) maximum negative face value of -0.78 is observed at Face-A at 60^0 of wind incidence. High negative face value (-0.77) is observed at Face-G at 90^0 wind incidences followed by Face-F and Face-L with decrement of about 4% at 180^0 wind incidence both due to flow separation from the faces at wide angles due to the orientation of faces to the direction of wind incidence. Face-E and Face-E are also experiencing high negative face values at 120^0 wind incidence angle due to the depressed orientation of faces. Highest face value is observed at Face-C (0.63) of 30^0 wind incidence angle closely followed by 150^0 and 180^0 of wind incidence with magnitudes of 0.57 (Face-G) and 0.60 (Face-H and Face-J) respectively mainly due to trapping of wind between the inner edges of the faces.



(a) RR (b) RF (c) FR (d) FF

Figure. 8 Pressure contour of Face-A of M-1 for interference condition at (a) RR, (b) RF, (c) FR and (d) FF

Interference Condition

Both the models have been studied for four interference conditions as, RR, RF, FR and FF orientation (as shown in Figure. 4) of the building with respect to twin their buildings located at the downstream of the wind stream. For all setups, the spacing between obstructing building models is fixed at 10% of the instrumented building model's height (60 mm).

The distribution of mean pressure coefficients is shown in Figure. 8 for four interference conditions. For RR condition it can be observed that the suction is generated for the entire Face-A of principal building with maximum suction of -0.29 at 1/3rd height of the model and minimum suction of -0.25 at 1/6th height of the model. From height 1/3rd to full height of model the pressure is very uniformly distributed, no influence of the interfering building model is observed from 1/6th to 1/3rd height of model. Uniform distribution can also be seen for FF interference condition throughout the height of the model with maximum suction of 0.56 at 1/6th height of the model from top and minimum suction of -0.40 at top and bottom of the model. However, from height 50mm to 200mm very less fluctuation can be observed. The uniformity of the pressure distribution along the height is due to the symmetrical plan shape as well as symmetrical orientation of both the models with respect to wind incidence. For three interference conditions except RR condition upwash is seen at 1/6th height of the model from top. The variation in the Cp for RF and FF conditions are high and ranges from -0.40 to -0.64 and -0.40 to -0.56 respectively. In both of RR and FF conditions due to the orientation of both the models, at Face-A of principal model wind is getting trapped between both the models and a shielding effect is created due the presence of interfering model with less variation of Cp (-0.24 to -0.28 for both the conditions). For interference conditions of RF, FR and FF least to no influence of interfering model is observed at 0mm to 100mm, 200mm to 300mm and 50mm to 200mm respectively.

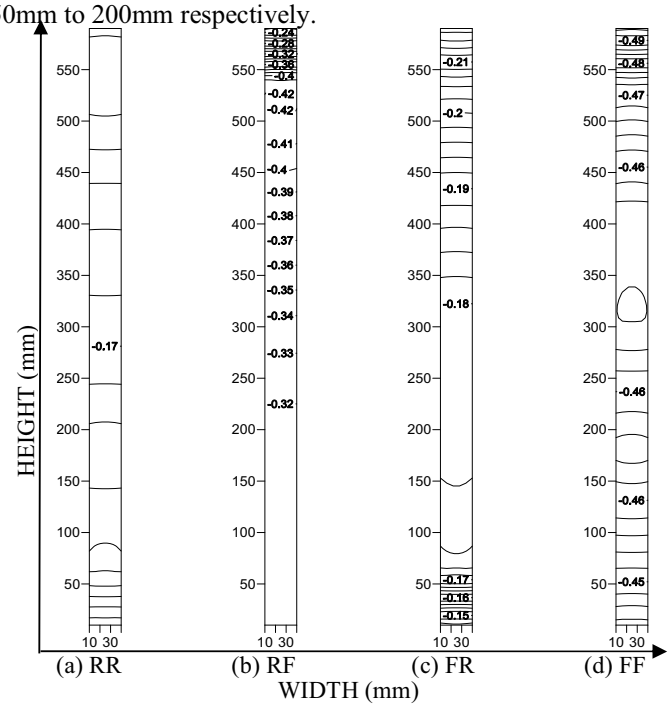
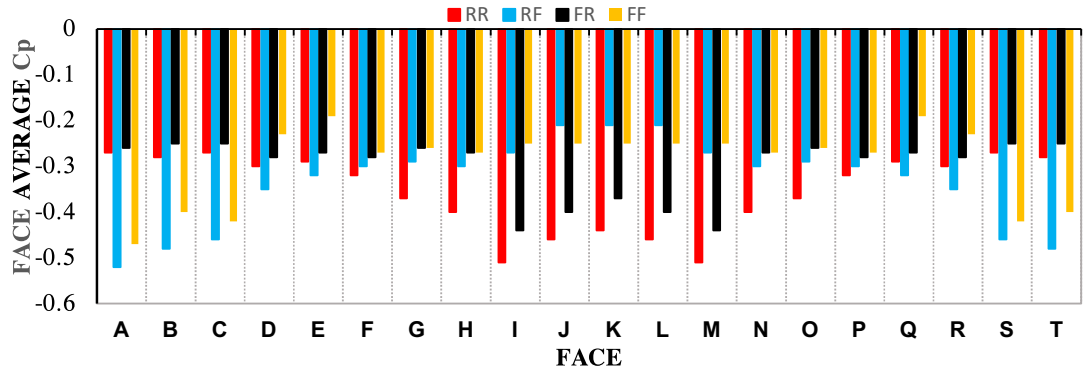


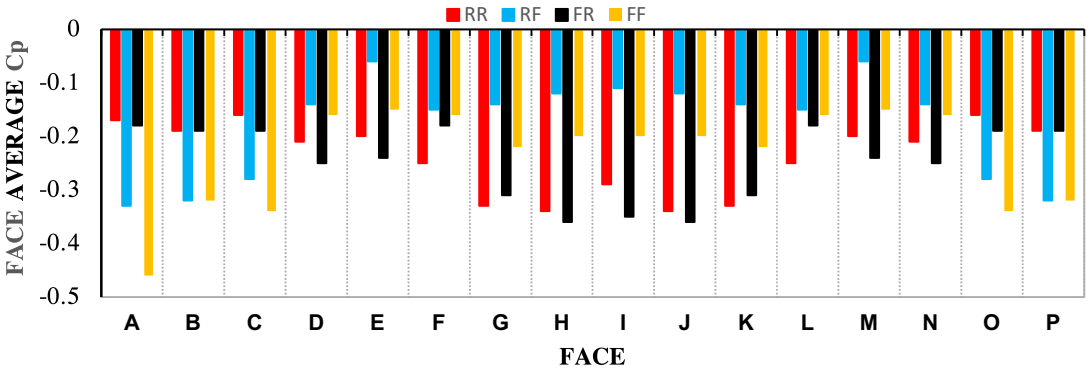
Figure. 9 Pressure contour of Face-A of M-2 for interference condition at (a) RR, (b) RF, (c) FR and (d) FF

Figure. 9 shows the distribution of mean wind pressure for four interference conditions. Face-A is least affected at RR because to the building's orientation, where the wind deviates symmetrically at greater angles on both sides after incidence on the obstructing model. Thus, very less suction is created at Face-A throughout the height. The maximum Cp (-0.185) is at the top and minimum Cp (-0.155) at the bottom of the model. From bottom to 1/10th height the wind is trapped thus rapid

fluctuation of the values can be observed. At RF and FR conditions influence of incident wind doesn't show any effect between bottom to 1/3rd and 1/6th to 1/2nd height of the model respectively. For RF condition the magnitude of pressure is varying from -0.32 to -0.42 and further reducing from -0.42 to -0.28 at 1/3rd to 7/6th and 7/6th to full height of the model respectively. A huge fluctuation at 7/6th to full height is observed due to the upwash phenomenon of the building as a result of presence of hindering building. In FR condition although the fluctuations are rapid but have very less change in the magnitude and thus, varying from -0.18 to -0.218 from 1/2nd to full height of the model. At bottom 1/10th of the height of model rapid change in the Cp can be observed as a result of entrapped wind at such height between both the models. However, no influence of interference of neighboring building model is observed from1/3rd to 1/2nd height of model in such condition. At FF orientation due to the symmetry of the stream of wind flow as well as the arrangement of the buildings the distribution of the Cp is also observed to be symmetrical horizontally as well as vertically. The magnitude of Cp is seen to increase from -0.44 to -0.48, fall from -0.48 to -0.45 at the model's bottom to 1/4, and climb from 1/4 to 1/2 height. Although a minor vortex can be seen at half the model's height, a similar distribution is shown at the upper half of the model. All interference cases exhibit upwash, with the exception of the RR orientation, which resembles M-1.



(a) Model-1



(b) Model-2

Figure. 10 Face Average Cp distribution al the periphery (Interference condition)

at (a) Model-1; (b) Model-2

At M-1 (Figure 10 (a)) the minimum magnitude of face value (-0.19) is observed at Face-E and Face-Q of FF condition whereas maximum magnitude of negative face value (-0.52) is observed at Face-A of BF condition. Due to the front orientation of the principal model, huge fluctuation is seen between the face values along the faces of the model. Due to maximum shielding of principal model of M-2 (Figure 10(b)) at BB condition least fluctuation between the face values among the

faces is seen. Whereas at BF condition maximum least magnitude of negative face value (-0.06) is seen at Face-E and Face-M. Whereas maximum magnitude of negative face value of -0.46 is observed Face-A of FF condition.

The blockage effect happens when one or more objects block the fluid flow's path from reaching the principal object. The existence and impact of an intruding structure are indicated by the interference factor on the original structure. This factor is being used to assess the fluctuation in pressure at a surface on the instrumented tall building. The interference factor is evaluated by, Kar and Dalui (2016), Zu and Lam (2018).

$$\text{Interference Factor} = \frac{\text{Pressure for selected point in interfering condition}}{\text{Pressure for selected point in isolated condition}}$$

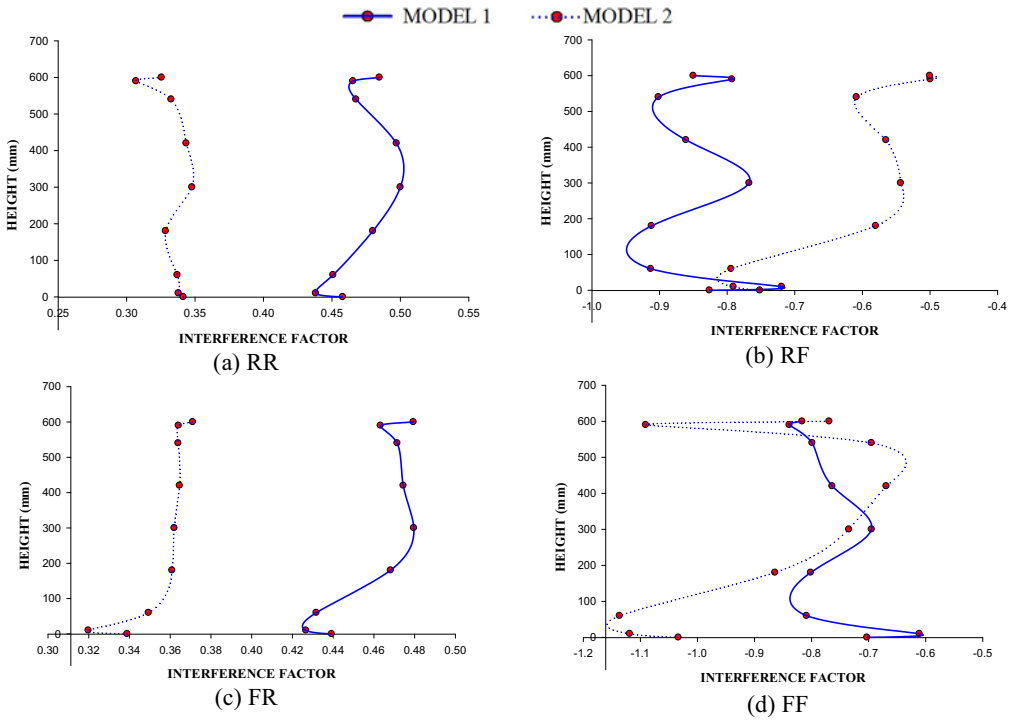
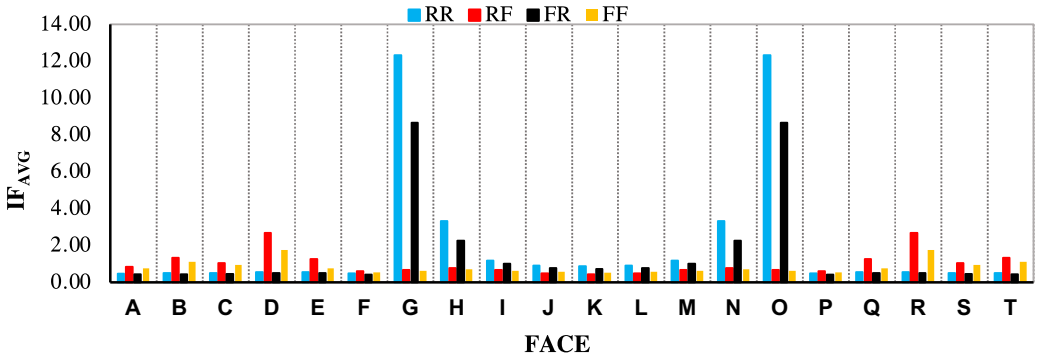


Figure.11 Interference Factor of Face-A along centerline (M-1 and M-2)

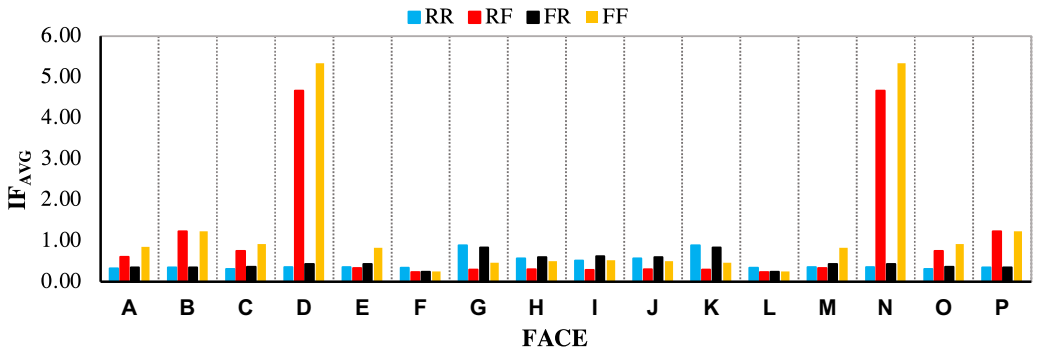
Figure. 11 shows comparison between the IF of M-1 to M-2 for similar interference conditions. For RR and FR interference condition variation of IF's along vertical centerline is almost similar for both the models with appreciable change in the magnitude. For RR condition the magnitude of IF for M-1 (maximum 0.50) is higher as compared to M-2 (maximum 0.35) which shows large reduction in IF for M-2 as compared to M-1. For FR a similar variation can be observed between magnitudes of IF's for M-1 (maximum 0.48) as compared to M-2 (maximum 0.37). At RF and FF interference condition a noticeable decrease to the Cp (IF<1) on the principal building is observed due to interference effect created by the presence and orientation of interfering building. In RF interference condition magnitude of IF M-1 (maximum -0.72) is higher as compared to M-2 (maximum -0.51) which shows large reduction in Cp for M-1 as compared to M-2. However, the variation of IF's for M-1 and M-2 at FF blockage condition have no similarity with large variation of Cp along the centerline of Face-A. The orientation of the interfering and the principal building models are playing vital role in the magnitudes of IF. The height between ½ to ⅔ of the height at the centreline of Face-A exhibits the largest influence of the obstructing model under all obstructing conditions; however, this height may vary depending on the topography.

The impact of the interfering building model is demonstrated in the following section by calculating the absolute magnitudes of Face Average Interference factors. The magnitudes of IF_{avg} are significantly more important than the magnitudes along the signs since the signs fail to accurately represent the IF and are entirely reliant on the pressure and/or suction magnitudes of C_p at isolated and interference building models.

$$\text{Face Average Interference Factor } (IF_{AVG}) = \frac{\text{Face Average } C_p \text{ for selected face in interfering condition}}{\text{Face Average } C_p \text{ for selected face in isolated condition}}$$



(a) M-1



(b) M-2

Figure 12. IF_{AVG} distribution along the periphery of (a) M-1 and (b) M-2

The wind is blowing at either 0^0 or 180^0 at both models under the interference condition, causing the models to behave symmetrically along the front Face-A. Figure 12. The highest magnitude (12.33) of IF_{AVG} is observed at symmetrical Faces G and O under the RR and FR interference circumstances of M-1. These high magnitudes are caused by the faces' depressed placements (refer to Figure 3) at the isolated wind situation at 180^0, which causes enormous oscillations to be noticed along the height of the faces. As a result, the faces' G and O face values abruptly change to 0.03 when compared to the surrounding faces. The neighboring symmetrical Faces H and N behave similarly to Faces G and O, respectively, for comparable causes. At RF and FF condition of M-2 model high the wind is flowing at 0^0 as similar to is to isolated wind condition. The symmetrical face pairs Face- D and N shows high magnitudes (4.67) of IF_{AVG} due to its depressed positions and large disturbances at isolated condition to 0^0 wind incidence and thus making the Face values very abrupt as compared to nearby face. In all the interference conditions the face value magnitudes doesn't show abrupt changes and are under wake region throughout.

4. Conclusions

This study emphasizes to find correlation between the effects of wind incidence on two distinct plan shape tall building models. The experiment has been carried out by creating Terrain Category

2 (IS: 875- Part-3 (2015), under a constant flow with the wind velocity of 10 m/sec. The plan area for Model-1 and Model-2 is kept same, as 40000mm². The effects of various wind incidences and also various blockage conditions between twin buildings has been studied separately at the nib Model-1 and Model-2 of i.e., Face-A through experimental setup within BLWT. A comparison study of mean pressure coefficients and interference factor along the vertical centreline between Model-1 and Model-2 has been carried out. For isolated conditions, wind angle is varied from 0° to 180° (at an interval of 30°) and for interference conditions, four different orientations between the twin buildings were used as RR, RF, FR, and FF. The conclusions drawn from the experimental results are as follows.

1. Experimental results of wind pressure coefficients of Model-1 and Model-2 shows good agreement with various wind incidence conditions with minute difference in the magnitudes of Cp. However, large difference is observed between both the models for 60° wind incidence condition where, the nature of pressure is changing throughout the vertical centerline of the models.
2. For both the model negative pressure is observed for all interference conditions due to presence of interfering model.
3. In Model-1 and Model-2, when the principal models are faced front irrespective of the orientation of interfering models, high suction is developed between both the twin models as compared to rear-facing principal models where, flow separation is taking place.
4. Among all the interference conditions the /Cp variation is seen between ½ to ⅔ of the height at the centerline of Face-A
5. At the interference condition the face average magnitudes at Model-1 is dominated by the orientation of interfering models; whereas that in Model-2 is dominated by the orientation of the principal model.
6. At all the conditions it is observed that the wind is getting trapped between ground and 10mm height of the model whereas, upwash is seen at 590mm height of the model.
7. On the basis of wind tunnel studies, it can be distinguished that the performance of Model-2 is better than Model-1 in both isolated as well as interference conditions for all wind incidences and orientations respectively.

References

[1] Alminhana WG, Braun LA and Loredo-Souza MA. A numerical-experimental investigation on the aerodynamic performance of CAARC building models with geometric modifications 2018 *J. Wind Engineering & Industrial* Aerodynamics Vol: **180** 34-48.
[2] Ahmed E, Bitsuamlak G and Damatty AE. Enhancing wind performance of tall buildings using corner aerodynamic optimization 2017 *Engineering Structures* Vol: **136** 133–148.
[3] Amin JA and Ahuja AK. Aerodynamic modifications to the shape of the buildings: a review of the state-of-the-art 2010 *Asian Journal of Civil Engineering (Building and Housing)* Vol. **11** no. 4 433-450.
[4] Amin JA and Ahuja AK. *W*ind-Induced Mean Interference Effects Between Two Closed Spaced Buildings 2012 *KSCE Journal of Civil Engineering* Vol: **16(1)** 119-131.
[5] AS/ NZS: 1170.2:2011. Structural design actions, Part 2: Wind actions. Australian/ New Zealand Standard, 2011.
[6] ASCE: 7-02. Minimum Design Loads for Buildings and Other Structures. 2002.
[7] Bairagi AK and Dalui SK. Optimization of interference effects on high rise buildings for different wind angles using CFD simulation 2014 *Electronic Journal of Structural Engineering* 2014 Vol: **14**.
[8] Bhattacharyya B, Dalui SK and Ahuja AK Wind Induced Pressure on 'E' Plan Shaped Tall Buildings 2014 *Jordan Journal of Civil Engineering* Volume **8** No. 2.
[9] Chakraborty S, Dalui SK and Ahuja AK *Experimental and Numerical Study of Surface Pressure on '+' Plan Shape Tall Building* 2013 *J. Civil Engineering* Vol: **8** 251-262.

[10] Cheng L, Kit M and Wong SY POD analysis of crosswind forces on a tall building with square and H-shaped cross sections 2015 *Wind and Structures, An International Journal* Vol: **21(1)** 63- 84. DOI: 10.12989/was.2015.21.1.063
[11] Chitra GS, Harikrishna P and Selvi RS Effects of upstream terrain characteristics on aerodynamic coefficients of structures 2017 *Archives of civil and mechanical engineering* **17** 776-785.
[12] Chuang WC and Spence SMJ A performance-based design framework for the integrated collapse and non-collapse assessment of wind excited buildings 2017 *Engineering Structures* Vol: **150** 746-758.
[13] Gomes M, Rodrigues A and Mendes P Experimental and numerical study of wind pressures on irregular-plan shapes 2005 *Journal of Wind Engineering and Industrial Aerodynamics* Volume **93** Issue **10** 741-756.
[14] Hangan H, Maryam R, Chowdhury J, Djordje R, Dan P, Julien L and Adrian C Novel techniques in wind engineering 2017 *J. Wind Engineering & Industrial Aerodynamics* Vol: **171** 12-33.
[15] Hassanli S, Hu G, Kwok KCS and Fletcher DF Utilizing cavity flow within double skin façade for wind energy harvesting in buildings 2017 *J. Wind Engineering and Industrial Aerodynamics* Vol: **167** 114-127.
[16] Hu G, Hassanli S, Kwok KCS and Tse KT Wind-induced responses of a tall building with a double-skin façade system 2017 *Journal of Wind Engineering and Industrial Aerodynamics* Vol: **68** 91-100.
[17] IS: 875- Part-3 (2015). Code of Practice for Design Loads (other than earthquake loads) for Buildings and Structures- Wind Loads. India.
[18] Kalehsar HE and Khodaie N 2017 Enhancing wind performance of tall buildings using corner aerodynamic optimization 2017 *Engineering & Structures* Vol: **136** 133-148.
[19] Kalehsar HE and Khodaie N Wind-induced vibration control of super-tall buildings using a new combined structural system 2018 *Journal of Wind Engineering & Industrial Aerodynamics* Vol: **172** 256-266.
[20] Kar R and Dalui SK Wind interference effect on an octagonal plan shaped tall building due to square plan shaped tall buildings 2016 *Int J Adv Struct Engg* Vol: **8** 73–86.
[21] Khanduri AC, Bedard C, and Stathopoulos T Modelling wind-induced interference effects using back propagation neural networks 1997 *Journal of Wind Engineering and Industrial Aerodynamics* Vol. **72** 71-79.
[22] Kheyari P. and Dalui SK Estimation of Wind Load on a Tall Building under Interference Effects: A Case Study 2015 *Jordan Journal of Civil Engineering* Volume **9** No. 1.
[23] Kim YC and Kanda J Wind pressures on tapered and set-back tall buildings 2013 Journal of Fluids and Structures Vol: **39** pp. 306–32.
[24] Li Y, Li. QS and Chen F Wind tunnel study of wind-induced torques on L-shaped tall buildings 2017 *J. Wind Engineering and Industrial Aerodynamics* Vol: **167** 41-50.
[25] Ming G Wind-resistant studies on tall buildings and structures 2010 *Science China Technological Sciences* Vol. **53** No.**10** 2630–2646. doi: 10.1007/s11431-010-4016-2
[26] Mukherjee S and Bairagi AK Interference effect on principal building due to setback tall building under wind excitation 2018 *SEC18: Proceedings of the 11th Structural Engineering Convention* Paper No. 20180306.
[27] Paul R and Dalui SK Wind effects on 'Z' plan-shaped tall building: A case study 2016 *International Journal of Advanced Structural Engineering* Vol: **8** 319–335.
[28] Santosh G and Amin JA Numerical simulation of Wind induced mean interference between two tall buildings 2017 *Journal of Materials and Engineering Structures* Vol: **4** 181–192.
[29] Silva FT, Takaaki K, Carlos P, Oscar GL and Jia C A review of computational fluid dynamics (CFD) simulations of the wind flow around buildings for urban wind energy exploitation 2018 *J. Wind Engineering & Industrial Aerodynamics* Vol: **180** 66-87.
[30] Solari G Wind Loading of Structures: Framework, Phenomena, Tools and Codification 2017 *Structures* Vol: **12** 265–285.

[31] Stathopoulos T and Zhou Y Numerical simulation of wind induced pressures on buildings of various geometries 1993 *J. Wind Eng. Ind. Aerod* Vol: **46–47** 419–430.
[32] Tse KT, Weerasuriya AU, Zhang X, Li SW and Kwok KCS Effects of twisted wind flows on wind conditions in passages between buildings 2017 *J. Wind Engineering and Industrial Aerodynamics* Vol: **167** 87-100.
[33] Xu X, Yang Q, Yoshida A and Tamura Y Characteristics of pedestrian-level wind around super-tall buildings with various configurations 2017 *J. Wind Engineering and Industrial Aerodynamics* Vol: **166** 61-73.
[34] Zheng C, Xie Y, Khan M, Wu Y and Liu J Wind-induced responses of tall buildings under combined aerodynamic control 2018 *Engineering Structures* Vol: **175** 86-100.
[35] Zu GB and Lam KM Across wind excitation mechanism for interference to twin tall buildings in tandem arrangement 2018 *Wind and Structures* Vol. **26** No. 6, 397-41.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

CDME-GAT: Context-Aware Depression Detection Using Multiembedding and Graph Attention Networks in Social Media Text

Minni Jain, *Member, IEEE*, Siddh Jain, Amita Jain, and Bhavuk Garg

***Abstract*—Depression, a prevalent mental health concern, requires timely identification and intervention. Automating the early stage identification of depression cues within social media text has become critically important. Existing methods for depression identification through text obtained from social media have shown promising results; however, these methods do not address the graphlike nature of social media data. To date, there has been little work addressing this problem by appropriately modeling social media data. This article aims to effectively harness the power of multiembedding techniques and graph attention networks (GATs) for depression identification. To overcome these limitations, the current study presents a novel methodology that combines multiple token-level embeddings, including BERT, RoBERTa, and DeBERTa, with GATs to leverage the graphlike nature of social media data and detect depression cues from such text. A dataset comprising 100 000 tweets has been curated using data from publicly available annotated tweet datasets. This dataset maintains a balanced distribution between depressive and nondepressive text samples. This was followed by a rigorous cleaning and preprocessing pipeline. Next, these data were transformed into a numeric feature matrix using multiple-word embeddings, which enabled the treatment of tweets as nodes in a graphlike structure. This graph was used for training a GAT model with multiple self-attentional layers, culminating in a linear layer for mapping traits to binary classes. The model presented in the current study achieves a precision of 97.2%, a recall of 96.4%, and an F1 score of 96.7% on a benchmark dataset.**

***Index Terms*—Depression detection, graph attention network (GAT), graph inference, mental health assessment, multiembedding.**

I. Introduction

DEPRESSION is a common mental disorder marked by loss of pleasure or interest in activities for long periods. As reported by the World Health Organization (WHO), an estimated 3.8% of the population experiences depression, including 5% of adults (4% among men and 6% among women), and 5.7% of adults older than 60 years; approximately 280 million people in the world have depression[1]; life events such as loss, childhood adversity, and unemployment contribute to the development of depression; and this has yielded repercussions spanning the social, economic, and health domains.

Depression among women is about 50% more common than among men. Worldwide, more than 10% of pregnant women and those who have just given birth experience depression [1]. Although there are known, effective treatments for mental disorders, more than 75% of people in low and middle-income countries receive no treatment [2]. Access to quality mental health care faces obstacles globally, including a lack of funding, a shortage of adequately trained healthcare professionals, and the social stigma attached to mental disorders. Current circumstances have raised the urgency of early depression identification and preemptive measures to extend support to those affected. In today's world of social media, people resort to sharing their lives on digital platforms. This data can be analyzed to find precious insights into the mental state of individuals. Thus, analyzing text available on social media has emerged as a promising approach for early depression detection and proactive mental health care.

The current study presents a comprehensive investigation into depression detection through textual analysis of social media data using graph attention networks (GATs). GATs, a class of neural networks, have showcased remarkable prowess in capturing dependencies intrinsic to graphlike structures [3]. They accurately encapsulate interactions among neighboring nodes, thus presenting an enriched contextual comprehension of the data. Contrasted against conventional models such as LSTM [4] and CNNs [5], proven to excel in sequential and gridlike formations, GATs wield a distinctive edge in handling the graphlike nature of social media data. By leveraging masked self-attentional layers, GATs can incorporate graph topology and capture relational information efficiently. This is particularly valuable for text classification tasks that require an adept understanding of the context and semantics of a given text.

Manuscript received 29 November 2023; revised 29 March 2024; accepted 13 May 2024. This work was supported by the IIT Mandi iHub & HCI Foundation under the Ministry of Science & Technology, Government of India. *(Corresponding author: Amita Jain.)*

Minni Jain is with the Department of Computer Science and Engineering, Delhi Technological University, New Delhi 110042, India (e-mail: minni.jain@dtu.ac.in).

Siddh Jain and Amita Jain are with the Department of Computer Science and Engineering, Netaji Subhas University of Technology, New Delhi 110078, India (e-mail: siddh.jain.ug21@nsut.ac.in; amita.jain@nsut.ac.in).

Bhavuk Garg is with the Department of Psychiatry, Lady Hardinge Medical College, New Delhi 110001, India (e-mail: dr.bhavukgarg@lhmc-hosp.gov.in).

Digital Object Identifier 10.1109/TCSS.2024.3402689

[1]Institute of Health Metrics and Evaluation. Global Health Data Exchange (GHDx). VizHub - GBD Results (healthdata.org)

2329-924X © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:51:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Beyond the model architecture, this article addresses critical aspects such as selecting an optimization algorithm and using multiple-word embeddings instead of just one for feature extraction. The current study proposes an ensemble method for the embeddings generated by pretrained language models (PLMs) such as BERT [6], RoBERTa [7], and DeBERTa [8] for formulating a feature matrix, which helps capture various aspects of the textual data.

Furthermore, the article presents a comprehensive text preprocessing pipeline to ensure dataset quality and enhance model performance. This pipeline spans data collection, text cleaning, stemming, and length-based filtering among other methods, collectively weaving a balanced and refined dataset. Thus, the biases and noise in the data are mitigated, ensuring the model's robustness and generalization capabilities.

Contributions:

1) The current study leverages an extensive dataset of around 100 000 social media posts, which helps broaden the scope of the research.
2) Introducing an efficient method for constructing feature matrix-based graphs, valuable for modeling complex textual relationships.
3) Using GATs for precise depression detection in social media text, capitalizing on attention mechanisms and graph-based representations to focus on important words.
4) The proposed approach represents an advancement over the current state-of-the-art methods in depression detection through textual sequences.

Further sections are as follows. Section II casts a spotlight on related work in the fields of depression detection and social media analysis. Section III details the proposed methodology, covering the dataset, preprocessing techniques, model architecture, and model training. Section IV includes experimental setup and evaluation of time complexity and outcomes, with result analysis and comparison with baselines following in Section V. Finally, Section VI concludes the study, highlighting the contributions, limitations, and pathways for forthcoming research works in this area.

II. Related Works

The increased use of social media in the last decade has paved the way for new opportunities to access and analyze user-generated content. However, the vast amount of data generated online and the sensitivity of mental-health conditions calls for leveraging natural language processing (NLP) to automate mental health analysis for early detection and prevention of depression and stress.

In the initial phases of research, statistical attributes were used, which were then integrated into established machine learning methods, as demonstrated by Castillo-Sánchez et al. [9], involved the adoption of well-known techniques such as SVM, random forest, and logistic regression, and employed to detect conditions pertinent to mental health.

Recent advancements in mental-health-related tasks have demonstrated notable progress by utilizing deep-learning models to capture semantic information [11], [12], [13], [14].

Furthermore, the integration of attention mechanisms underscores how they effectively make models to focus on the most significant part of the input [13], [15], [16], [17], [18].

PLMs such as BERT, detailed by Haque et al. [19], and RoBERTa, exemplified by Murarka et al. [20], play a pivotal role in identifying various mental health concerns. These models harness their adept contextual modeling capabilities to provide substantial insights. In a noteworthy study, Ji et al. [22] curated a comprehensive dataset from social media platforms and utilized pretrained BERT to post-train a domain-specific model called MentalBERT, which demonstrated improved performance across multiple depression detection datasets.

Graph-based neural networks, such as graph convolutional networks (GCNs) [23] and GATs [3], excel in social media text analysis due to their capacity to capture relationships, context, and interactions within interconnected data [24].

Ensemble learning, a technique combining multiple models, enhances predictive performance while reducing variance, bias, and noise in data analysis [25], [26], [27].

Table I summarizes the contributions of some notable studies in this field.

Previous studies have predominantly focused on detecting depression in social media text using individual PLMs or deep-learning methods such as LSTMs and GCNs [23]. However, these studies have often been constrained by small datasets, limiting their applicability considering the vast volume of data generated online every second. In contrast, the current study leverages a comprehensive dataset of 100 000 tweets, significantly broadening the research scope. Additionally, we employ an ensemble of multiple PLMs to construct the feature matrix for the texts, facilitating the incorporation of richer semantic information and enhancing model performance. The current study introduces an innovative and effective approach for graph construction from the feature matrix, enabling accurate modeling of the inherent graphlike nature of social media data. By utilizing GATs [3] for data modeling and depression detection, our study surpasses the limitations of existing methods and offers a more robust solution.

III. Proposed Methodology

The primary objective of the proposed research work is to detect depression, from public posts collected from social media, particularly Twitter. We have a sample set $\{U_i, V_i\}$ in which $i \in [1,N]$, where each of the posts U_i is labeled with a label $V_i \in \{0, 1\}$, and N is the size of the dataset. The goal is to search for potential indicators of depression in every post U_i and accordingly label it as depressive (indicated by "1") or nondepressive (indicated by "0"). Therefore, the challenge at hand can be defined as a binary text classification problem. Here, a function (P) with model parameters γ is developed

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:51:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
COMPARATIVE ANALYSIS OF RELATED WORKS

Name and Year	Method(s)	Contributions
R. Chiong et al. (2021) [10]	Linear SVMs, Logistic Regression, Decision Trees, and Multilayer Perceptron	Explored the effectiveness of various machine learning methods in detecting depression signs from social media posts.
Kim et al. (2020) [11]	XGBoost and CNN	Investigated whether specific Reddit posts of the user can be classified as relevant types of mental disorder.
Ghosh and Anwar (2021) [12]	LSTM	Aimed to predict depression and estimate its intensity in social media users.
Sekulic and Strube (2019) [13]	GRU with attention mechanisms	The study employs a hierarchical attention network (HAN), for depression detection from Reddit posts.
Yadav et al. (2022) [14]	Bidirectional Gated Recurrent Unit (BiGRU) network	Presented an approach to identify depression based on linguistic material from patient interviews using a fully coupled network.
Wang et al. (2021) [15]	C-Attention Network	Used latent features to detect suicide attempts from social media post data of individuals.
Yao et al. (2021) [16]	LSTM with attention mechanisms	Automatic extraction of depressive symptoms from online depression communities.
Ren et al. (2021) [17]	BiLSTM with attention mechanisms (BiLSTM_Att)	Pursued an investigation using Reddit data, for mental health analysis.
K. Yang et al. (2022) [18]	Knowledge–aware and contrastive network (KC-NET)	This approach employed GRUs and a knowledge–aware metallization module with dot-product attention to model mental states and detect depression.
Tejaswini et al. (2024) [21]	Fasttext Convolution Neural Network with Long Short-Term Memory (FCL)	Combined FastText embedding, CNN architecture for global information extraction, and LSTM architecture to extract local features with dependencies, to detect depression through text classification.
Naseem et al. (2022) [24]	Text Graph Convolutional Network (TextGCN), BiLSTMs, and attention mechanisms	Detected depression severity levels on Reddit using ordinal classification.
Adarsh et al. (2023) [25]	SVM + KNN	Introduced an ensemble model with intrinsic explainability and noisy label correction to distinguish depression symptoms and suicidal ideation, addressing participation imbalances through a one-shot decision approach.
Firoz et al. (2023) [26]	CNN-LSTM + BiLSTM	Proposed a hybrid model with fine-tuned hyperparameters, evaluated with 189 clinical and Wizard-of-OZ interviews.
Jyoti et al. (2024) [27]	Gaussian Naïve Bayes classifier + Bi-LSTM	Introduced an ensemble model to extract contextual semantics from text using Part-of-Speech (POS) tagging and Word2Vec.

to classify each post (text sample) U_i into the correct label V_i, shown as follows:

$$\text{argmax } \gamma \prod_{i=1}^{N} P(V_i \mid U_i, \gamma).$$

Prior to delving into the subsequent sections, it is important to address some preliminary topics pertinent to the subject matter of this research article.

A. Preliminary

NLP involves using computers to perform various tasks related to understanding and analyzing human language. There are several state-of-the-art models capable of processing natural language text. These models can analyze large volumes of complex textual data to extract valuable insights.

1) Context-Aware Word Embeddings: The development of context-aware word embeddings has significantly advanced the field of NLP. This evolution is exemplified by the progression from BERT [6] to RoBERTa [7] and ultimately to DeBERTa [8], each introducing novel techniques to enhance language understanding and performance across various NLP tasks.

Introduced by Google AI researchers in 2018, bidirectional encoder representations from transformers (BERTs) revolutionized NLP by capturing contextual information from both left and right contexts. Trained on extensive corpora such as the Toronto Book Corpus and Wikipedia, BERT employs bidirectional transformers and pretraining objectives such as masked language modeling and next-sentence prediction. Its ability to capture nuanced relationships between words has led to remarkable performance in tasks such as text classification and question answering.

Following BERT's success, Facebook AI introduced RoBERTa in 2019 with a focus on refining pretraining methodologies. By increasing model size, utilizing dynamic masking, and removing the next-sentence prediction task, RoBERTa achieves improved representations and better generalization in downstream tasks. Furthermore, RoBERTa adopts a larger byte pair encoding (BPE) vocabulary, enabling it to handle a broader range of subword units and rare words more effectively.

In 2020, Microsoft Research introduced DeBERTa, which builds upon RoBERTa's advancements and further refines word order modeling and sentence coherence. By incorporating disentangled attention and an enhanced mask decoder, DeBERTa allows for more granular analysis of token relationships, capturing complex linguistic structures more accurately. Additionally, DeBERTa incorporates relative position bias in self-attention and parameter sharing across layers, resulting in superior performance, particularly in scenarios with limited training data or resource constraints.

2) GATs: GATs, a subset of graph neural networks (GNNs), have garnered attention for their adeptness in generating robust representations for NLP tasks. Introduced by Veličković et al. [3], GATs employ masked self-attentional layers to effectively weigh information from neighboring nodes in a graph. This enables GATs to learn intricate patterns and relationships within graph-structured data, proving valuable for tasks involving graphlike data, such as social media networks. Unlike traditional neural networks, which assume fixed-size vector inputs and struggle with graph-structured data, GNNs excel in differentiating, identifying, and establishing relationships between words and phrases.

An overview of the proposed methodology has been provided in Fig. 1. The method begins with (a) data collection and dataset structuring followed by (b) data preprocessing. Then, (c) the input text is subjected to multiple text encoders, which facilitate

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:51:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

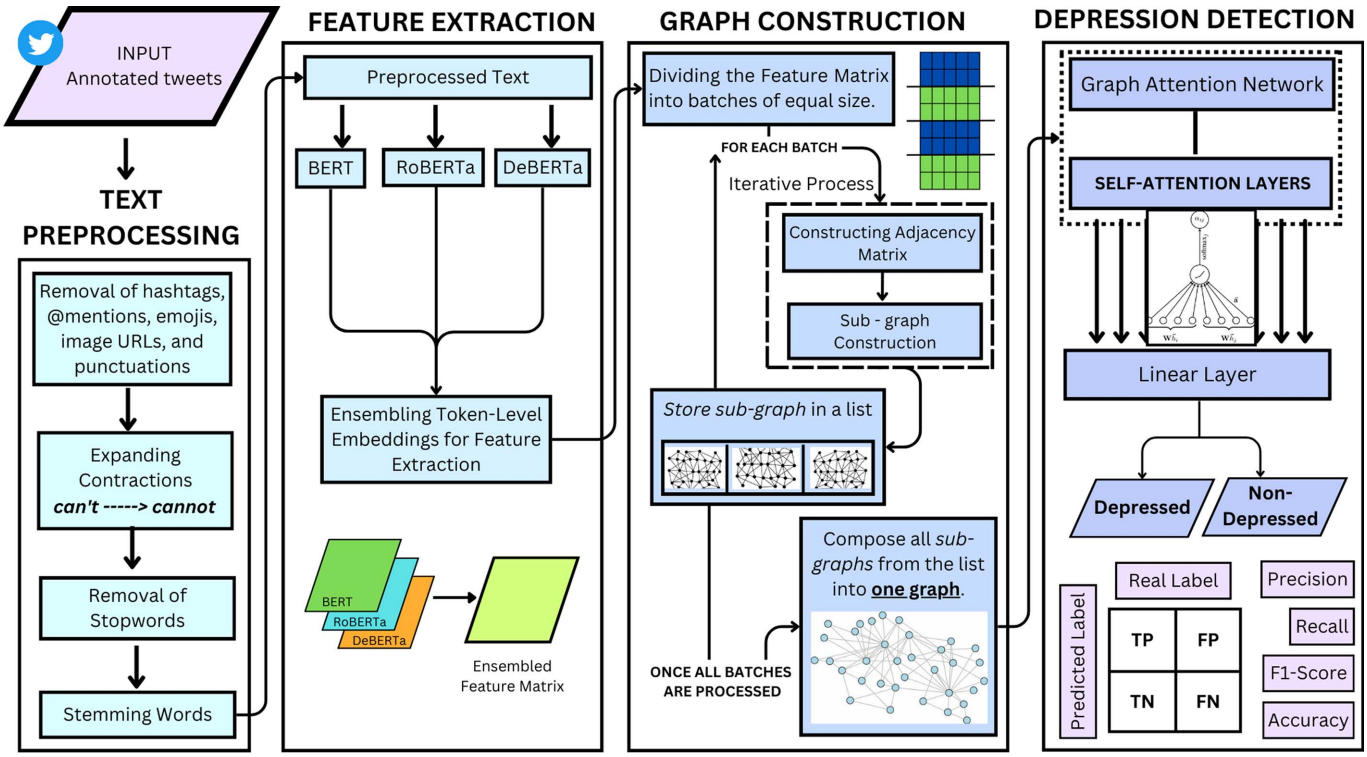


Fig. 1. Overall architecture of proposed methodology.

the generation of word embeddings for deeper language comprehension. The (d) embeddings thus obtained are stacked to create a 3-D feature representation which is then (e) adjusted back to 2-D by fusion along the z-axis. Further, (f) the feature matrix is transformed into a graph, (g) which is then used to train the classification model. Further sections discuss these steps in detail.

A. Dataset Preparation

1) Data Collection: The data largely include Twitter posts. Additionally, some open-source datasets such as Sentiment 140, which contains tweets classified as positive, negative, and neutral, have also been incorporated. While it is always intended to ensure the credibility and quality of the data sources, it is important to note that the verification process mainly pertains to the dataset itself rather than individual posts or statements within the dataset. Table II provides a comprehensive overview of the datasets that contribute to the final dataset utilized in this study, along with detailed descriptions of their contents.

2) Dataset Structuring: For constructing the dataset, following steps were taken. First, it was ensured that the resulting dataset has equal number of posts of both labels, i.e., for a dataset of size N, $N/2$ posts should be the ones labeled as "depressive" and $N/2$ posts should be labeled as "nondepressive."

Consider the set of depressed posts $X = \{X_1, X_2, X_3, \ldots, X_n\}$ and the set of nondepressed posts $Y = \{Y_1, Y_2, Y_3, \ldots, Y_n\}$, where $n = N/2$. Afterwards, these N posts are arranged alternatively to construct the final dataset. This approach ensures an equal number of posts of each class belonging to training, testing, and validation sets cutout from this dataset, thus avoiding

TABLE II
DATASET DESCRIPTION

Dataset Name	Source Datasets	Dataset Size	Content Details
Depressive Tweets Mega Dataset	I am diagnosed with depression [38]	58	Tweets with the theme that the author has been diagnosed with depression.
	I am fighting depression [38]	80	Tweets with the theme that the author is fighting depression.
	I suffer from depression [38]	554	Tweets with the theme that the author is suffering from depression.
	Tweetdata.txt [39]	160 454	Text file containing a large repository of tweets that sound depressed.
	Depressive_ tweets_ processed [40]	2323	Comprises tweets that have been labeled as depressed.
Random Tweets Dataset	Sentiment140 [41]	200 000	1 600 000 annotated tweets extracted using the twitter API. 200 000 tweets were randomly selected from this dataset.
Final Dataset	Depressive Tweets Mega Dataset	50 000	50 000 tweets selected at random from this dataset containing 163 469 tweets labeled as depressed.
	Random Tweets Dataset	50 000	50 000 tweets which did not suggest depression, were selected at random from this dataset

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:51:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

any kind of class imbalance at any point in the dataset. Hence, the final dataset of posts $U = \{U_1, U_2, U_3, \ldots, U_N\}$, is constructed such that

$$U = \{X_1, \quad Y_1, \quad X_2, \quad Y_2, \quad X_3, \quad Y_3, \quad \ldots, X_n, \quad Y_n\}.$$

B. Text Preprocessing

The dataset comprises text collected from social media posts, which often contain noise in the form of slang words, emojis, hashtags, and incorrect punctuation. To address this, the text undergoes cleaning through several steps. Initially, hashtags, URLs, image URLs, emojis, and @mentions to other user accounts are removed. Subsequently, punctuation marks are eliminated to further reduce noise. Contractions are then expanded to their original base form words, enhancing understanding of sentiments by the encoder and model. Finally, each post is tokenized into words using NLTK's word tokenizer. Thus, the ith post in the data $U^i = \{U^i_1, U^i_2, \ldots, U^i_{Ni}\}$, where U^i_j is the jth token of the ith post, and N_i denotes the number of tokens in U^i. Now, all the *stop words* are removed from ith post, thus, reducing the ith post to $\hat{U}^i = \{\hat{U}^i_1, \hat{U}^i_2, \ldots, \hat{U}^i_{Ki}\}$, where $\hat{U}^i_j$ is the jth token and K_i denotes the number of tokens in $\hat{U}^i$, such that $K_i \leq N_i$.

After this, each of the tokens in the post $\hat{U}^i$ is reduced to its base form using PorterStemmer. For example, the words "running" and "ran" are both reduced to "run," and the words "depression" and "depressing" are both reduced to "*depress*." Finally, all elements in the processed ith post, U^i, are rejoined, such that the final input is of the form $U^i = \{U^i_1; U^i_2; \ldots; U^i_{Ni}\}$, where ";" denotes concatenation and N_i denotes the number of tokens in the ith post. This concludes the text-cleaning process and M cleaned posts are obtained.

C. Context-Aware Text Encoder

PLMs have transformed NLP by excelling in comprehending and generating humanlike text, leveraging extensive training on diverse textual data. They offer numerous advantages in text modeling, including contextual understanding facilitating interactions across lengthy sequences. Additionally, PLMs integrate semantic information and pretraining knowledge, enhancing sentence-level representations. Their adaptability and efficiency in handling language nuances, along with the ability to capture intricate linguistic structures, further improve the modeling process. However, for more fine-grained representations, it is better to consider not one but multiple such sentence encoders, each with its specific properties, to capture the semantics of the text from various viewpoints. Hence, the current study follows this approach of using multiple context-aware text encoders, BERT [6], RoBERTa [7], and DeBERTa [8] to obtain token-level embeddings.

The input representations are obtained through each encoder using E layers of the transformer with the preprocessed input U_i. This process can be denoted as $X_i = \text{encoder}_k(U_i, E)$ where $X_i \in R^N \times D_h$ denotes the outputs of the Eth layer of the kth text encoder, encoder_k. Here, R is the set of real numbers, N denotes the length of the sequence, and D_h is the hidden dimension of the kth text encoder.

D. Ensemble of Token-Level Embeddings for Feature Extraction

The embeddings for every encoder are obtained as a matrix of dimensions $M \times D_h$ where M is the number of posts in the dataset and D_h is the hidden dimension of the encoder. For uniformity, the value of D_h for all three encoders is ensured as equal. For k encoders, we obtain k matrices of dimensions $M \times D_h$ and stack them along a third axis to obtain a new 3-D matrix of size $M \times D_h \times k$. This matrix is then flattened by compressing the values across the newly introduced dimension by average pooling, as shown in Fig. 2. This process yields a matrix of context-aware token-level embeddings of dimensions $M \times D_h$, where M is the number of posts in the dataset and D_h is the hidden dimension of the encoders. The matrix thus obtained is the feature matrix.

E. Graphical Modeling of Textual Data through Batch Processing

This section emphasizes the creation of a graph from extracted features. The primary goal is to depict relationships between data points (posts) as edges in a graph, with nodes representing the data points. To construct the graph, the first step involves creating the adjacency matrix, a square matrix used to represent a finite graph. Each element within the matrix indicates whether pairs of vertices in the graph are adjacent.

For constructing the adjacency matrix, we obtain a dot product of the feature matrix with its transpose. Thus, for a dataset having M posts, the dimensions of the adjacency matrix will be $M \times M$. In the resulting adjacency matrix, a row is in the format $X^i = \{W^i_1, W^i_2, W^i_3, \ldots, W^i_M\}$, where X^i denotes ith post and W^i_j denotes the weight of the edge between X^i and X^j. The weight indicates the strength or significance of the relationship between the posts. Due to the potentially large size of the dataset, directly constructing a graph from the adjacency matrix may not be feasible due to space and time constraints. Hence, the current study adopts a novel approach to overcome this obstacle and construct graphs efficiently. Algorithm 1 guides the processing of the feature matrix in batches, with each batch containing C posts. For each batch, an adjacency matrix is constructed, leading to the formation of a distinct graph known as a "*subgraph*."

A detailed explanation of the construction of subgraphs has been provided as follows. Consider a feature matrix of dimensions $M \times D_h$, where M is the number of posts and D_h is the hidden dimension of the encoder. Every batch of the feature matrix of dimensions $C \times D_h$ is processed to construct a subgraph. So, for every batch of C posts, the batch adjacency matrix is obtained as the dot product of the batch feature matrix and its transpose, thus resulting in a batch adjacency matrix of dimensions $C \times C$. Each row of this matrix is of the form, $X^i = \{W^i_1, W^i_2, W^i_3, \ldots, W^i_C\}$, where X^i denotes ith post of that batch and W^i_j represents the weight of the edge between X^i and X^j. A graph constructed from this batch adjacency matrix using the *NetworkX* library is the subgraph. We iterate over the feature matrix

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:51:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

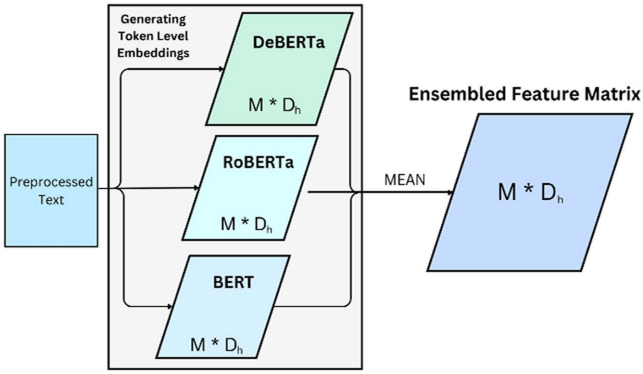


Fig. 2. Ensemble of token-level embeddings.

in batches, constructing subgraphs for each batch. This process continues until all data points have been processed, resulting in *M/C* subgraphs. Each subgraph obtained after processing a batch is stored in a listlike data structure for future access. This allows for efficient memory management and access to individual subgraphs as needed.

In the concluding step, all subgraphs are combined into a unified graph. The composition involves a straightforward union of both node sets and edge sets. It is worth noting that the node sets of the input graphs do not necessarily have to be disjoint. Additionally, attributes of the graph, its edges, and nodes are propagated to this union graph. Hence, we obtain a graph of M vertices and $M \times M$ edges.

F. Attention-Based Supervised Learning Model for Depression Detection

After constructing the graph, the "*edge index*," which is a tensor detailing the edge connectivity information of the graph, is obtained. It is usually a tensor of shape *(2, num edges)*, where each column contains a pair of node indices (source and target) that define an edge in the graph. Here, "*num edges*" stands for the number of edges in the graph. This tensor specifies how nodes are connected.

Next, the dataset splits into training, validation, and testing datasets. To create these sets, "masks" are used, which are binary arrays that specify which data points belong to each set. These masks help to ensure control over how data split and prevent data leakage (i.e., ensuring that data used for validation and testing are not inadvertently included in the training set). Also, separate tensors are used for the node feature matrix and the target labels, which are used to create the "data" object for representing graph data.

1) Model Overview: GATs [3] are employed for training due to their robust representations and learning capabilities, particularly notable in applications such as recommendation systems, NLP, and healthcare. Given the graph-natured data from social media, these models are particularly well suited for such datasets. In GCNs [23], every neighboring node has the same importance. This should not be the case since some nodes are more essential than others. GATs offer a solution to this problem by leveraging masked self-attentional layers. To consider the importance of each neighbor, an attention mechanism assigns a weighting factor to every connection. In this way, the model addresses several key challenges of spectral-based graph neural networks simultaneously.

The current study uses multiple self-attentional layers instead of just one which helps the increase model capacity to capture intricate graph patterns and relationships, enabling the GAT to learn hierarchical representations of the graph and promote better modeling of extracted features. As proposed by Veličković et al. [3], through their article, GATs [3], the model working is detailed as follows. Each attention layer in the graph has an input set of node features $v = \{v_1, v_2, \ldots, v_N\}$ where $v_i \varepsilon R^G$, where N denotes the number of nodes and G is the number of features in each node. The output is a new set of node features $v' = \{v'_1, v'_2, \ldots, v'_N\}$, $v'_i \varepsilon R^{G'}$, where G' is the updated cardinality value.

The process is initialized with each node undergoing a linear transformation, enhancing its expressive capabilities to elevate input features into higher level representations. Subsequently, the model engages in self-attention across nodes, employing a shared attentional mechanism denoted as a: $R^{F'} \times R^{F'} \rightarrow R$ to calculate attention coefficients, $c_{ij} = a\,(Wh_i, Wh_j)$, that indicate the importance of node j's features to node i. Hence, the model "allows every node to attend on every other node, dropping all the structural information." Notably, the computation of c_{ij} is confined to all the nodes j within the neighborhood of node i, achieved through the implementation of masked attention to incorporate the graph structure into the mechanism.

2) Model Training: The training process starts by initializing the GAT [3] model class. The model is initialized with the number of nodes in the input graph, which is determined by the shape of the feature matrix, and the number of classes, which is set to *2* for binary classification tasks. The Adam [32] optimizer is used as the optimization algorithm. We define the cross-entropy loss [34] as the loss function to measure the difference between predicted and true class labels.

The model is trained by iterating through multiple epochs, as described in Algorithm 1. In every epoch, the model is first set to training mode to enable dropout and batch normalization layers to operate differently compared to the evaluation phase. Then, the gradients of the optimizer are reset to zero before computing gradients in each iteration. Next, the input features and graph edge indices are passed to the GAT [3] model to obtain node embeddings (h) and class predictions (z). The loss in each iteration is computed by comparing the predicted class labels (z) with the true labels *(data.y)* only for the training nodes, as indicated by the "train_mask." The training and validation accuracies are calculated by comparing the model's predictions with the ground truth labels, considering only the respective subsets of nodes ("train_mask" and "val_mask").

Finally, the loss gradients are computed with respect to the model's parameters, and these weights are updated using the optimizer's update rule. Upon completing all epochs, the trained GAT model is returned for prediction on new data or further analysis. Model performance is assessed on the test data to analyze the effects of hyperparameters such as learning rate, number of epochs, and self-attentional layers on the results.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:51:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Algorithm 1 CDME-GAT

Input: Tweets (posts from twitter)
Output: Classification of the tweet (depressive or nondepressive)

1. Read the dataset and perform text-preprocessing.
2. Obtain token-level word embeddings using BERT, RoBERTa, and DeBERTa.
   $X_i = encoder_k(U_i, E),\ X_i \in R^N \times D_h$
3. Ensemble the word embeddings to form the feature matrix. (F_m)
4. Divide the feature matrix into batches and create a list.
5. For each batch B_i of dimensions $C \times D_h$, where C is the number of posts per batch and D_h is the hidden dimension of the encoders:
   I. Construct the adjacency matrix (A_m) of dimensions $C \times C$:
   $A_m = (F_m).(F_m{}^T)$
   II. Construct a subgraph (G_{sub}) and store it in the graph list (L_G)
   $G_{sub} = Graph(A_m)$
   $L_G.append(G_{sub})$
6. Compose all the *subgraphs* in the list to form a single graph (G) representing token-level relationships.
   $G = compose_all(L_G)$
7. Obtain the edge index (E_{ix}) of the graph.
8. Split the dataset into train, validation, and test datasets.
9. Create train, validation, and test masks as *"train_mask,"* *"val_mask,"* and *"test_mask,"* respectively.
10. Create separate tensors x and y for storing the features associated with each node in the graph and labels associated with each node.
    $x = tensor(F_m, dtype = float)$
    $y = tensor(df['label'], dtype = long)$
11. Create a "Data" object for representing the graph data.
    $data = Data(x = x, edge_index = E_{ix}, y = y)$
12. Define a custom accuracy function which calculates accuracy as a percentage of the correctly predicted instances.
13. Initialize the GAT model class.
    $model = GAT(num_nodes = F_m.shape[1], num_classes = 2)$
14. Initialize the optimizer (O) and the loss function (L_F).
    $O = Adam(model.parameters(), lr = 0.01)$
    $L_F = CrossEntropyLoss()$
15. For each training epoch
    I. Set the model in training mode.
    II. Zero out the gradients of the optimizer.
    III. Compute node embeddings (h) and predictions (z) using the GAT model.
    IV. Calculate the loss using specified criterion, considering only the training nodes:
    V. Calculate the training and validation accuracies using the custom accuracy function defined in step 11:
    VI. Backpropagate the gradients through the GAT model.
    VII. Update the model's parameters using the optimizer
16. Return the trained GAT model.

IV. Experimentation Setup, Time Complexity and Results

This section details the experimental setup for the presented work. Various metrics such as test accuracy, precision, recall and F1 score are considered to evaluate the model performance. These metrics help to get a deeper idea of how well the model performs under different experimental settings. The current study considers the macroaverage F1 score for the ablation study and performance comparison with other baselines, which calculates metrics for each label and finds their unweighted mean. Since there is no class imbalance in the dataset considered, the macroaverage and weighted average F1 scores would be equal here.

A. Experimental Configurations

In this section, we have discussed the experimentation performed to observe the effect of various hyperparameters on model performance. The text-preprocessing methods, as detailed in Section III-B, involved several essential steps to ensure dataset quality and enhance the performance of our text classification model. A balanced, cleaned dataset of 100 000 texts was prepared through this preprocessing pipeline, focusing on relevant content and equal representation of depressed and nondepressed samples. 60% of the dataset was used for training the model, and the remainder was equally split for validation and testing of the model performance.

Experiments with different embeddings and ensembles, listed in Tables III and IV, informed the feature matrix selection. For the proposed model, an ensemble of the matrices obtained from BERT [6], RoBERTa [7], and DeBERTa [8] was used. We have experimented with two models, GCNs [23] and GATs [3].

For optimization of the GCN [23] model, adaptive moment estimation with weight decay (AdamW) [33] was used, which is an optimization algorithm that builds upon the Adam [32] optimizer. It helps stabilize the optimization process and can lead to better generalization of performance. However, for the GAT [3] model, which uses self-attentional layers, Adam proved to be better at optimization than AdamW. Hence, it was concluded, that the effectiveness of the optimization algorithm depends on the specific dataset, problem, and architecture. Final model configuration, detailed in Section IV-C, includes a single GAT layer with multiple masked self-attentional layers and a Linear layer projecting outcomes into depressive and nondepressive categories.

B. Time Complexity

This section elaborates on the assessment of the asymptotic time complexity of the proposed approach. The approach is segmented into five phases, and the time complexity of each phase is determined based on Algorithm 1.

1) Text Preprocessing: The time complexity for text preprocessing is determined as $O(N * (n * (1 + 8 * k) + m))$, where N is the total number of tweets, n is the average length of tweets, k is the average word length, and m is the average number of words per tweet.

2) Feature Extraction: The feature extraction phase is characterized by a time complexity of $O(N * (n + D_h))$, where N is

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:51:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III
CDME-GCN—EXPERIMENTAL AND ABLATION STUDY RESULTS

Exp. No.	Embeddings and Experimental Settings	Test Acc.	P	R	F1
1	Word2Vec + Glove + FastText (AdamW lr = 0.01, weight decay = 0.04, 800 epochs)	91.95	93.38	90.29	91.80
2	Glove + FastText (AdamW lr = 0.01, weight decay = 0.04, 800 epochs)	92.17	94.30	89.77	91.98
3	**BERT + RoBERTa +DeBERTa (AdamW lr = 0.01, weight decay = 0.04, 500 epochs)**	**93.99**	**95.05**	**92.97**	**94.01**
4	BERT + RoBERTa +DeBERTa (AdamW lr = 0.01, weight decay = 0.04, 800 epochs)	93.84	94.79	92.84	93.89

Note: Exp. No.: experiment number, Test Acc.: test accuracy, P: precision, R: recall, F1: F1 score, lr: learning rate. The bold entries correspond to the experimental configurations delivering superior results for the models CDME-GCN and CDME-GAT, respectively.

the total number of tweets, n is the average length of tweets, and D_h is the hidden dimension of the encoders.

3) Graph Construction: The time complexity of the graph construction phase is expressed as $O(N * c^2)$, where N is the total number of tweets and c is the size of the submatrix or batch size.

4) Calculation of Edge-Indices: The time complexity for calculating edge indices is $O(E)$, where E is the number of edges in the graph.

5) Model Training: Model training exhibits a time complexity of $O(E * D_h)$, where E is the number of edges in the graph and D_h is the hidden dimension of the encoder which is also the number of features representing each node of the graph.

Combining the individual time complexities of all phases of the proposed approach, we get the final asymptotic time complexity as

$$O(E * D_h + N * (n * (1+8 * k) + m + c^2)).$$

C. Experimental Results Analysis and Ablation Study

Details of our extensive work and results across various experimental configurations are presented in Tables III and IV. We first trained a GCN [23] model with Word2Vec [35], Glove [36], and FastText [37] embeddings, all of which have $D_h = 512$. The model achieved satisfactory results on the concerned evaluation metrics. The optimizer used here is AdamW [33], and the number of training epochs is 800. The experimentation was done over a range of hyperparameter values. Despite excluding Word2Vec due to poor performance, negligible improvement was observed. With GCN [23] based model architectures, it was found that training required a higher number of epochs (in the range of 500–1000) to achieve good results on the given data. Later, we changed the embedding set and used BERT [6], RoBERTa [7], and DeBERTa [8], each of which has $D_h = 768$. It was found that the results improved for every metric by 2%–3% as we experimented with different learning rates and weight decay values. The experimental configuration yielding the best overall performance for the CDME-GCN model is highlighted in Table III.

The CDME-GCN model configuration utilized the AdamW optimizer with a weight decay of 0.04. For feature representation, BERT, RoBERTa, and DeBERTa were employed. The learning rate was set to 0.01, and the model was trained for 500 epochs. This configuration achieved an accuracy of 93.99%, precision of 95.05%, recall of 92.97%, and an F1 score of 94.01% on the test dataset.

As detailed in Table IV, extensive experimentation was conducted on the GAT [3] model, with different embedding sets across a varied range of the number of self-attentional layers. During the experimentation phase, a few points became clear. First, the GAT model outperformed the GCN model, requiring fewer training epochs for optimal results. Second, BERT, RoBERTa, and DeBERTa embeddings exhibited superior performance in representing textual data compared to Word2Vec, Glove, and FastText.

Adam proved more effective in optimizing the CDME-GAT model than AdamW, which performed well with CDME-GCN. Increasing the number of attention heads initially improved results, but eventually plateaued or declined. The optimal configuration, highlighted in Table IV, utilized 64 self-attentional layers. It employed BERT, RoBERTa, and DeBERTa embeddings, with a learning rate of 0.01 and 250 epochs. This configuration achieved an accuracy of 94.58%, precision of 96.67%, recall of 93.29%, and an F1 score of 94.36% on the test dataset. Figs. 3 and 4 illustrate the visualizations of the results obtained from the ablation study.

D. Validation of Proposed Approach Through Statistical Significance

Model validation ensures the fulfillment of the study's objectives by comparing the model's output to the output of the data-generation process. The model performance is validated through a z-test conducted on two random samples extracted from the dataset employed in the present study. In sample 1, 5120 posts were considered (N_1), of which 4655 (X_1) were correctly classified. The proportion of correctly classified tweets here is $p_1 = 0.9092$. Sample 2 comprised 4000 posts (N_2), of which 3635 (X_2) were correctly classified. So, the proportion of correctly classified tweets is $p_2 = 0.90875$.

The following set of null (H_0) and alternate (H_a) hypotheses was developed to test the model's validity.

(H_0): proportion of correctly classified tweets in both samples is the same, $p_1 = p_2$.

(H_a): proportion of correctly classified tweets differs in both samples, $p_1 \neq p_2$.

The value of the pooled proportion is calculated as

$$P = \frac{X_1 + X_2}{N_1 + N_2} = \frac{4655 + 3635}{5120 + 4000} = 0.909.$$

The z-statistic for a two-tailed test can be calculated as

$$z = \frac{p_1 - p_2}{\sqrt{P(1-P)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} = \frac{0.9092 - 0.90875}{\sqrt{0.909 \times (0.091)\left(\frac{1}{5120} + \frac{1}{4000}\right)}} = 0.074.$$

As $|z| = 0.074 \leq z_c = 1.96$, we fail to reject the null hypothesis. Therefore, at the $\alpha = 0.05$ significance level, there is

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:51:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE IV
CDME-GAT—EXPERIMENTAL AND ABLATION STUDY RESULTS

Exp. No.	Attention Heads	Embeddings and Experimental Settings	Test Acc.	P	R	F1
1	4 Heads	Word2Vec + Glove +FastText (Adam lr = 0.01, 500 epochs)	92.98	94.53	91.23	92.85
2	8 Heads	Word2Vec + Glove +FastText (Adam lr = 0.01, 500 epochs)	93.18	94.38	91.82	93.08
3	16 Heads	Word2Vec + Glove +FastText (Adam lr = 0.01, 500 epochs)	93.28	94.24	92.19	93.20
4	32 Heads	Word2Vec + Glove +FastText (Adam lr = 0.01, 160 epochs)	93.81	94.91	92.22	93.71
5	32 Heads	Word2Vec + Glove +FastText (Adam lr = 0.01, 500 epochs)	92.68	92.71	92.65	92.68
6	64 Heads	Word2Vec + Glove +FastText (Adam lr = 0.01, 160 epochs)	93.62	94.01	93.18	93.59
7	64 Heads	Word2Vec + Glove +FastText (Adam lr = 0.01, 500 epochs)	93.02	93.39	92.58	92.98
8	4 Heads	Glove + FastText (Adam lr = 0.01, 160 epochs)	92.64	94.39	90.66	92.48
9	8 Heads	Glove + FastText (Adam lr = 0.01, 160 epochs)	93.33	94.90	91.57	93.21
10	16 Heads	Glove + FastText (Adam lr = 0.01, 160 epochs)	93.38	94.62	92.04	93.29
11	32 Heads	Glove + FastText (Adam lr = 0.01, 160 epochs)	93.57	95.32	91.63	93.44
12	64 Heads	Glove + FastText (Adam lr = 0.01, 160 epochs)	93.53	94.64	92.29	93.45
13	4 Heads	BERT + RoBERTa + DeBERTa (Adam lr = 0.01, 250 epochs)	94.15	96.05	92.08	94.02
14	8 Heads	BERT + RoBERTa +DeBERTa (Adam lr = 0.01, 250 epochs)	94.01	94.98	92.92	93.94
15	16 Heads	BERT + RoBERTa +DeBERTa (Adam lr = 0.01, 250 epochs)	93.76	95.50	91.84	93.63
16	32 Heads	BERT + RoBERTa + DeBERTa (Adam lr = 0.01, 250 epochs)	93.97	95.43	93.12	93.93
17	64 Heads	BERT + RoBERTa + DeBERTa (Adam lr = 0.01, 200 epochs)	94.26	96.12	92.68	94.11
18	**64 Heads**	**BERT + RoBERTa + DeBERTa (Adam lr = 0.01, 250 epochs)**	**94.58**	**96.67**	**93.29**	**94.36**
19	64 Heads	BERT + RoBERTa + DeBERTa (Adam lr = 0.01, 300 epochs)	94.12	95.92	93.02	94.04

Note: Exp. No.: experiment number, Test Acc.: test accuracy, P: precision, R: recall, F1: F1 score, lr: learning rate. The bold entries correspond to the experimental configurations delivering superior results for the models CDME-GCN and CDME-GAT, respectively.

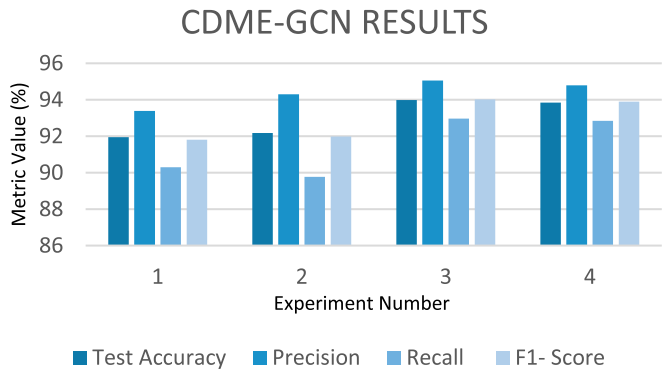


Fig. 3. Visualizing results of CDME-GCN model.

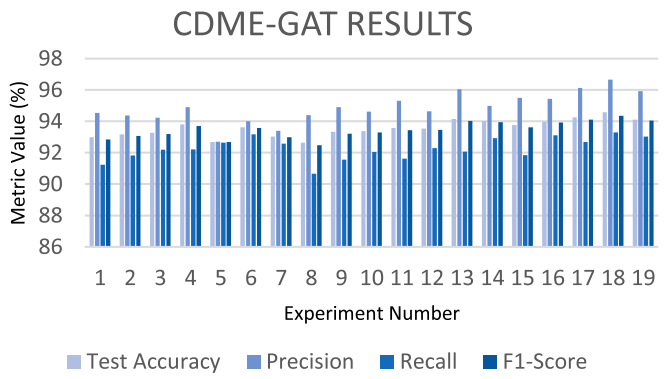


Fig. 4. Visualizing results of CDME-GAT model.

inadequate evidence to suggest a difference in the population proportions p_1 and p_2. Hence, the accuracy of the proposed model remains consistent across various sample sizes of the same dataset. Consequently, the findings of the CDME-GAT model for depression detection are deemed reliable.

V. PERFORMANCE COMPARISON

In this section, we delve into the comprehensive experimental performance of the baseline models and the proposed methods on the Depression_Mixed dataset [31]. This dataset, focused on weakly supervised depression detection, encompasses subreddits from Reddit. Each post within the dataset is categorized as either indicative of depression or not, providing a collection of monologues that detail the speaker's mental state. The evaluation is conducted across precision, recall, and F1 score metrics for a thorough comparison.

A. Overall Results

The overall experiment results of our model and the baselines are listed in Table V. An overview of this table shows that PLM-based models such as BERT [6] and RoBERTa [7] perform better than others with over 90% score across all metrics. MentalRoBERTa [22], which has implicitly infused domain-specific knowledge in this case, performs better than both BERT [6] and RoBERTa [7] but cannot match up to the performance of the proposed model. This shows the advantage of capturing the text semantics through various viewpoints by employing multiple PLMs.

The outcomes obtained from other baselines, such as logistic regression combined with features [29], and deep-learning methods such as CNN [28] and GRU [30], fall short of satisfaction. Notably, BiLSTM_Att [17], which incorporates attention mechanisms alongside Bidirectional LSTM, demonstrates superior performance, underscoring the effectiveness of attention mechanisms in addressing these challenges. KC-NET [18], leveraging mental state knowledge extracted from COMET and infusing it via GRUs, delivers robust results with an impressive 95.5% precision, 95.3% recall, and 95.4% F1 score. A more recent method introduced an ensemble model combining the Gaussian Naïve Bayes classifier and Bi-LSTM to extract contextual semantics from text using POS tagging and Word2Vec [27]. However, the results of this approach fell short of expectations.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:51:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE V
MODEL PERFORMANCE COMPARISON

Methods	Description	P	R	F
BiLSTM + Naïve Bayes [27]	Ensemble model using Naïve Bayes and BiLSTM along with POS tagging and Word2Vec	75.0	70.2	68.2
GRU [30]	GRU-based model with stacked layers for binary classification. The hidden states of the final time-step are used for decoding.	84.4	84.4	84.4
CNN [28]	A three-channel convolution as the encoder. Post-level embeddings are obtained using max-pooling and features of optimized word embeddings are concatenated for decoding.	85.2	85.1	85.1
LR+Features [29]	Logistic Regression (LR) combining selected linguistic features	89.0	92.0	89.0
BERT [6]	Initialized by the pre-trained weights of BERT-base.	91.4	91.4	91.4
BiLSTM_ Att [17]	Attention mechanism is used on the output of a bidirectional LSTM network. The attended hidden states are used for decoding.	90.4	95.0	92.6
RoBERTa [7]	Initialized with the pre-trained weights of RoBERTa-base.	93.2	92.4	92.9
MentalRoBERTa [22]	Initialized from the pre-trained weights of MentalRoBERTa.	93.4	93.0	93.3
KC-Net [18]	Extracted mental state knowledge from COMET and infused it using GRUs to explicitly model the mental states of the speaker.	95.5	95.3	95.4
CDME-GCN	Ensembles multiple word embeddings and models the features on a Graph Convolutional Network.	95.3	94.3	94.8
CDME-GAT	Ensembles multiple word embeddings and models the features on a GAT.	**97.2**	**96.4**	**96.7**

Note: P: precision, R: recall, F1: F1 score.

CDME-GCN attains a noteworthy performance on the Depression_Mixed [31] dataset with 95.3% precision, 94.3% recall, and 94.8% F1 score. However, the proposed method, CDME-GAT, establishes new state-of-the-art results by surpassing the baseline PLM-based methods with a remarkable 3%–4% improvement and outperforming current state-of-the-art methods by 1%–2% across all metrics. The proposed CDME-GAT model achieves outstanding metrics of 97.2% precision, 96.4% recall, and 96.7% F1 score on this dataset, underscoring its effectiveness in handling longer sequences and intricate textual content.

VI. CONCLUSION

The current study presents CDME-GAT, a multiembedding GAT for identifying texts that suggest depression. First, the feature matrix is extracted from the token-level embeddings representing the preprocessed tweets. Then, a graph constructed from these features is used to train a GAT with multiple masked self-attentional layers. After extensive experimentation, it is concluded that the model performs best with 64 self-attentional layers, beyond which the results tend to either saturate or drop depending upon other hyperparameters such as the learning rate and optimization algorithm. The obtained traits are mapped to a binary classification by a linear layer.

Our model demonstrates improved results over current state-of-the-art methods for depression detection on a benchmark dataset. Despite this success, there remains room for enhancing model performance. The ensemble of multiple embeddings, a critical aspect of our method, could benefit from further refinement by using different pooling techniques and studying their impact on the model performance. Additionally, advancements in the model architecture specifically tailored for depression detection could yield improvements. A more comprehensive approach to model social media data with improved representations can provide superior classification results. As the field of NLP continues to evolve, we anticipate the emergence of more domain-specific PLMs that could further elevate the quality of results.

While the present study focuses on detecting depression through textual content sourced from social media, it is crucial to acknowledge that online data exists in various forms, including images, videos, audio, and more. Efforts must be made, to develop methods capable of processing multimodal data to extract valuable information and model it effectively for depression detection. Additionally, it is essential to explore methods to classify the intents or causes of depression, moving beyond mere identification for a more in-depth analysis of a patient's situation and improved treatment strategies. Notably, mental health care often lacks the attention it deserves in many places globally. We aspire that the findings of this study contribute to early stage depression detection, potentially benefiting as many individuals as possible.

REFERENCES

[1] C. Woody, A. J. Ferrari, D. Siskind, H. Whiteford, and M. Harris, "A systematic review and meta-regression of the prevalence and incidence of perinatal depression," *J. Affect. Disorders*, vol. 219, pp. 86–92, Sep. 2017, doi: 10.1016/j.jad.2017.05.003.

[2] S. Evans-Lacko et al., "Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the WHO World Mental Health (WMH) surveys," *Psychol. Med.*, vol. 48, no. 9, pp. 1560–1571, Nov. 2017, doi: 10.1017/s0033291717003336.

[3] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lió, and Y. Bengio, "Graph attention networks," 2018, *arXiv:1710.10903*.

[4] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM—A tutorial into long short-term memory recurrent neural networks," 2019, *arXiv.1909.09586*.

[5] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/tnnls.2021.3084827.

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., vol. 1: Long and Short Papers*, 2018, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[7] Y. Liu et al., "ROBERTA: A robustly optimized BERT pretraining approach," 2019, *arxiv.1907.11692*.

[8] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," 2020, *arxiv.2006.03654*.

[9] G. Castillo-Sánchez et al., "Suicide risk assessment using machine learning and social networks: A scoping review," *J. Med. Syst.*, vol. 44, no. 12, p. 205, Nov. 2020, doi: 10.1007/s10916-020-01669-5.

[10] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104499, doi: 10.1016/j.compbiomed.2021.104499.

[11] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Sci. Rep.*, vol. 10, no. 1, p. 11846, Jul. 2020, doi: 10.1038/s41598-020-68764-y.

[12] S. Ghosh and T. Anwar, "Depression intensity estimation via social media: A deep learning approach," *IEEE Trans. Comput. Social Syst.*,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:51:56 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

vol. 8, no. 6, pp. 1465–1474, Dec. 2021, doi: 10.1109/tcss.2021.3084154.
[13] I. Sekulić and M. Strube, "Adapting deep learning methods for mental health prediction on social media," in *Proc. 5th Workshop Noisy User-Generated Text (W-NUT)*, Hong Kong, China, 2020, pp. 322–327.
[14] U. Yadav and A. K. Sharma, "A novel automated depression detection technique using text transcript," *Int. J. Imag. Syst. Technol.*, vol. 33, no. 1, pp. 108–122, Jul. 2022, doi: 10.1002/ima.22793.
[15] N. Wang et al., "Learning models for suicide prediction from social media posts," in *Proc. 7th Workshop Comput. Linguistics Clin. Psychol. Improving Access*, 2021, pp. 87–92. Accessed: Mar. 30, 2024. [Online]. Available: https://aclanthology.org/2021.clpsych-1.9
[16] X. Yao, G. Yu, J. Tang, and J. Zhang, "Extracting depressive symptoms and their associations from an online depression community," *Comput. Human Behav.*, vol. 120, Jul. 2021, Art. no. 106734, doi: 10.1016/j.chb.2021.106734.
[17] L. Ren, H. Lin, B. Xu, S. Zhang, L. Yang, and S. Sun, "Depression detection on Reddit with an emotion-based attention network: Algorithm development and validation," *JMIR Med. Inform.*, vol. 9, no. 7, Jul. 2021, Art. no. e28754, doi: 10.2196/28754.
[18] K. Yang, T. Zhang, and S. Ananiadou, "A mental state knowledge–aware and contrastive network for early stress and depression detection on social media," *Inf. Process. Manage.*, vol. 59, no. 4, Jul. 2022, Art. no. 102961, doi: 10.1016/j.ipm.2022.102961.
[19] F. Haque, R. U. Nur, S. A. Jahan, Z. Mahmud, and F. M. Shah, "A transformer based approach to detect suicidal ideation using pre-trained language models," in *Proc. 23rd Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dhaka, Bangladesh, 2020, pp. 1–5, doi: 10.1109/ICCIT51783.2020.9392692.
[20] A. Murarka, B. Radhakrishnan, and S. Ravichandran, "Classification of mental illnesses on social media using RoBERTa," in *Proc. 12th Int. Workshop Health Text Mining Inf. Anal.*, 2021, pp. 59–68. [Online]. Available: https://www.aclweb.org/anthology/2021.louhi-1.7/
[21] V. Tejaswini, K. S. Babu, and B. Sahoo, "Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model," *ACM Trans. Asian Low-Resour. Lang. Inf. Process*, vol. 23, no. 1, pp. 1–20, Jan. 2024, doi: 10.1145/3569580.
[22] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "MentalBERT: Publicly available pretrained language models for mental healthcare," in *Proc. 13th Lang. Resour. Eval. Conf.*, 2021, pp. 7184–7190.
[23] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.
[24] U. Naseem, A. G. Dunn, J. M. Kim, and M. Khushi, "Early identification of depression severity levels on Reddit using ordinal classification," in *Proc. ACM Web Conf., 2022*, pp. 2563–2572, doi: 10.1145/3485447.3512128.
[25] V. Adarsh, P. Kumar, V. Lavanya, and G. R. Gangadharan, "Fair and explainable depression detection in social media," *Inf. Process. Manage.*, vol. 60, no. 1, Jan. 2023, Art. no. 103168, doi: 10.1016/j.ipm.2022.103168.
[26] N. Firoz, O. G. Beresteneva, A. S. Vladimirovich, M. S. Tahsin, and F. Tafannum, "Automated text-based depression detection using hybrid ConvLSTM and Bi-LSTM model," in *Proc. 3rd Int. Conf. Artif. Intell. Smart Energy (ICAIS)*, 2023, pp. 734–740, doi: 10.1109/ICAIS56108.2023.10073683.
[27] J. Singh, I. Mangotra, M. Jain, and A. Jain, "A hybrid approach for depression detection using word embedding, Naive Bayes and Bi-LSTM models," in *Proc. Adv. Data-Driven Comput. Intell. Syst.*, 2024, pp. 295–307.
[28] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep learning for depression detection of Twitter users," in *Proc. 5th Workshop Comput. Linguistics Clin. Psychol. Keyboard Clinic*, 2018, pp. 88–97.
[29] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, Jan. 2019, doi: 10.1109/access.2019.2909180.
[30] F. Sadeque, D. Xu, and S. Bethard, "UARizona at the CLEF ERISK 2017 pilot task: Linear and recurrent models for early depression detection," *PubMed*, vol. 1866, Sep. 2017. Accessed: Mar. 30, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29075167
[31] I. Pirina and Ç. Çöltekin, "Identifying depression on Reddit: The effect of training data," in *Proc. EMNLP Workshop SMM4H 3rd Social Media Mining Health Appl. Workshop Shared Task*, 2018, pp. 9–12.
[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv.1412.6980*.
[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[34] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2018, pp. 8792–8802.
[35] T. Mikolov, K. Chen, G. S. Corrado, and J. M. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[36] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.
[37] A. Joulin, É. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, vol. 2: Short Papers*, 2017, pp. 427–431.
[38] *Depression Tweets*. Kaggle, 2021. Accessed: Mar. 30, 2024. [Online]. Available: https://www.kaggle.com/datasets/imen12/depression-tweets/data
[39] *Twitter Dataset Based on Depressive Words*. Kaggle, 2021. Accessed: Mar. 30, 2024. [Online]. Available: https://www.kaggle.com/datasets/saurabhshahane/twitter-dataset-based-on-depressive-words/
[40] *depressive_tweets_processed*. Kaggle, 2021. Accessed: Mar. 30, 2024. [Online]. Available: https://www.kaggle.com/datasets/dongphilyoo/depressive-tweets-processed/
[41] A. Go, R. Bhayani, and L. Huang, *Twitter Sentiment Classification Using Distant Supervision*," CS224N Project Report, Stanford, 2009.

Minni Jain (Member, IEEE) received the Ph.D. degree in computer science and engineering from Delhi Technological University, New Delhi, India, with a dissertation on natural language processing using soft computing, in 2023. She serves as an Assistant Professor with Delhi Technological University. Her research focuses on natural language processing, mental health care, social media analysis, and artificial intelligence. She has authored numerous research papers published in prestigious journals and conferences, including IEEE TRANSACTIONS, ACM Transactions, Springer, and Wiley.

Siddh Jain is currently working toward the B.Tech. degree in computer science and engineering (big data analytics) with Netaji Subhas University of Technology, New Delhi, India.

Amita Jain received the Ph.D. degree in focusing on fuzzy based information retrieval: An NLP approach from Jawaharlal Nehru University, New Delhi, India. She is an Assistant Professor with Netaji Subhas University of Technology, New Delhi, India. With a prolific academic record, she has authored 120 research papers in esteemed international journals and conferences such as ACM Transactions, IEEE, Elsevier, and Springer. She has successfully supervised five Ph.D. students, currently guiding three others. Additionally, she has secured a research project on developing a Chatbot Counselling Application for depression disorder, funded by the Ministry of Science and Technology, India.

Bhavuk Garg received the M.B.B.S. degree from the prestigious Maulana Azad Medical College, New Delhi, India, and the M.D. degree in psychiatry from the Institute of Human Behaviour and Allied Sciences (IHBAS), New Delhi, India, where he also completed Senior Residency and later became Locum Consultant.

He is currently working as an Associate Professor in psychiatry with Lady Hardinge Medical College, New Delhi, India. He has 14 years of experience in psychiatry. He is actively involved in the fields of community psychiatry, child psychiatry, and addiction psychiatry.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:51:56 UTC from IEEE Xplore. Restrictions apply.

Check for updates

Research Article

Received: 24 February 2024 Revised: 31 May 2024 Published online in Wiley Online Library:

(wileyonlinelibrary.com) DOI 10.1002/jctb.7698

Concentration of betanin from model beetroot extracts by using nanofiltration: parameter estimation and sensitivity analysis

Ashwani Kumar Tiwari and Manish Jain*

Abstract

BACKGROUND: Nowadays, research focusing on natural compounds is getting more importance across the globe. Recently, natural bioactive compounds have gained the attention of researchers as a consequence of their exemplary physicochemical properties and exceptional utility in the food and pharma industries. However, separating and concentrating these bioactive compounds from their origin sources still poses a considerable challenge to their profitable commercial usage. The current study focuses on the application of a facile nanofiltration method for recovering betanin from model beetroot extract solutions using a polyamide nanofiltration membrane. Furthermore, the three-parameter Spiegler–Kedem model was used to determine the transport parameters of the membrane and theoretically predict the performance of the membrane. Additionally, variance-based sensitivity analysis methods were deployed to study the sensitivity of different operating and membrane transport parameters.

RESULT: The results indicate that the membrane exhibited the highest rejection of 99% and permeate flux of 3.14 10^{-5} m^3 m^{-2} s^{-1} at 0.6 MPa pressure and 600 mL min^{-1} flow rate. The data obtained through the Spiegler–Kedem model were coherent with the experimental observations. Simulations on sensitivity analysis revealed that specific hydraulic permeability and cross-membrane pressure majorly influence permeate flux and rejection.

CONCLUSION: This study showed that nanofiltration with 150 Da cut-off membranes effectively concentrated betanin with 99% rejection. Higher cross-membrane pressure and higher feed flow rates were preferable operating conditions for higher rejections. Variance-based sensitivity analysis showed that cross-membrane pressure and hydraulic permeability are the most influential parameters affecting permeate flux and rejection.
© 2024 Society of Chemical Industry (SCI).

Supporting information may be found in the online version of this article.

Keywords: Betanin; nanofiltration; phenolic compound; NF process modeling; sensitivity analysis

INTRODUCTION

Phenolic compounds are one of the most abundant subsidiary metabolites in plants and are identified by single or more aromatic rings that contain single or multiple hydroxyl groups.[1-3] These compounds offer many human-health-related functional benefits. Phenolics exhibit physiological benefits such as anti-aging, anti-infective, anti-inflammatory and antiproliferative properties.[4-7] Betanin is one such phenolic compound. It is a water-soluble vacuolar chromo-alkaloid found in *Beta vulgaris* L, also known as beetroot and chukandar.[8] Betanin finds applications in various industries, including the food industry as a colorant and the pharmaceutical industry as an active ingredient in medicines. It is also used for coloring meat substitutes, gelatine, desserts, sausages and dairy products.

The recovery of phenolic bioactive compounds such as betanin from their natural sources is tedious owing to the high enzyme activity in most foods and plants. Therefore, the choice of extraction method must be made with the greatest caution to prevent the chemical change of the target molecules.[9] The standard method for separating and concentrating phenolic compounds is liquid–liquid extraction followed by evaporation or crystallization. Liquid–liquid extraction requires organic solvents such as chloroform, *n*-hexane, ethanol, propanol, ethyl acetate, and acetone.[10,11] These solvent-based processes are very costly, and the phenolic extracts may contain residual solvents, making them unfit or undesirable for food applications.[1]

Nanofiltration (NF) seems to be a promising technology for concentrating phenolic compounds owing to certain benefits, such as no involvement of phase transition, low energy consumption, no use of organic solvents and low waste disposal. Moreover, other biological molecules such as sugars and proteins can be easily separated from phenolic solutions, and concentration and purification steps can be combined in one step, reducing operational costs and time. Membrane-based processes can run continuously

* *Correspondence to: M Jain, Department of Applied Chemistry, Delhi Technological University, New Delhi 110042, India. E-mail: manishjain@dtu.ac.in*

Department of Applied Chemistry, Delhi Technological University, New Delhi, India

and are easy to scale up to the commercial level.[12] The application of NF to concentrate phenolic compounds is a relatively new concept and is still being investigated. However, only a few studies are available on using NF for the concentration of bioactive compounds.[13-20]

Arboleda Mejia *et al*.[13] investigated the rejection percentage for total phenolic content (TPC) and proanthocyanidins and found 73% and 92% rejection using four NF membranes (CA series: laboratory-prepared by phase inversion and one commercial NF90). Tonova *et al*.[14] also separated the phenolic compounds from Eurasian water milfoil hydrolysate using NP030 (MWCO of 500 Da). Giordana *et al*.[15] have concentrated the bioactive compounds from strawberry juice using a polyvinylidene difluoride NF membrane with a molar mass cut-off ranging between 150 and 300 Da. They found that the NF of natura juice has retention values >95% for the TPC. In another similar work, Carmela Conidi *et al*.[17] concentrated the phenolic compounds from orange press liquor using four different membranes: NF-70 (polyamide+polysulfone), NF-200 (poly-piperazine amide), N30F (polyethersulfone) and NFPES10 (polyethersulfone) having MWCO of 180, 300, 400 and 1000 Da, respectively. The investigations revealed that all of these NF membranes showed a high average rejection percentage towards anthocyanins (89.2–95.9%) and flavonoids (70–95.4%). In other similar works, Mello *et al*.[20] concentrated the flavonoids and phenolic compounds in aqueous and ethanolic propolis extract using NF90 membrane (polyamide and polysulfone). They found that the aqueous solution of propolis retained almost 99% flavonoids and 84% phenolic compounds. However, the retention values were reduced to 90% for flavonoids and 53% for phenolic compounds during the filtration of ethanolic solutions.

In previous studies,[13-19,21,22] the authors have only concentrated the phenolic compounds, and a detailed investigation of the membrane transport phenomenon using mathematical models was not reported. Besides this, no literature has been found on the concentration of betanin using the NF membrane. The objective of the present study is to investigate the performance of the NF membrane for the concentration of betanin and predict the transport parameters. Accurate estimation of these parameters is crucial to optimize the NF process, design the scale-up setups and predict the viability of the process in commercial-scale operations. The Spiegler–Kedem model was used to describe the transport phenomenon and validate the data. Rigorous research was also conducted to analyze the effects of various operating conditions and transport parameters through sensitivity analysis.

MATERIALS AND METHODS

Materials

Laboratory-grade betanin (molecular weight 550.47 g mol^{-1}) was bought from TCI Chemicals, Adambakkam, India. The reverse osmosis (RO) water was collected from the RO water purification machine available at Delhi Technological University, Delhi. Flat sheet NF membranes (HFT-NF 150 and HFT-NF 300) were supplied by Permionics Pvt. Ltd, Gorwa Vadodara, India.

Experimental setup and procedure

A schematic diagram depicting the laboratory-scale experimental setup prepared for NF is presented in Fig. 1. The setup contained a test cell of stainless steel for mounting the NF membrane with an active membrane area of 0.007856 m^2. The other specifications of the membrane test cell were supplied by the manufacturer and

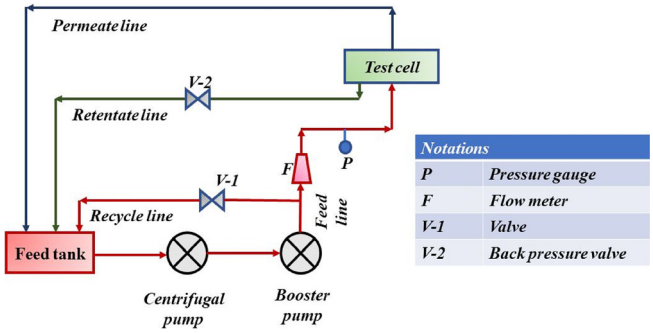


Figure 1. Block diagram of the prepared NF setup.

are mentioned in Table 1. A ¼ horsepower centrifugal pump followed by a booster pump (flow rate 100 GPD, maximum pressure 1.2 MPa) was employed for transporting the feed solution from the feed tank (capacity 10 L) to the membrane test cell. The setup also comprised a needle valve and a back pressure valve at the bypass and feed lines, which can be adjusted to control the feed stream flow rate and cross-membrane pressure. A flowmeter (rotameter of flow rate ranging from 100 to 1200 mL min^{-1}) and pressure gauge (glycerine filled 0 to 1.6 MPa) were employed to measure the feed flow rate and to maintain the hydraulic pressure, respectively. After filtration, separate tanks were used to collect the retentate and the permeate stream from the membrane test cell setup. The permeate flow rates and fluxes were determined by measuring the collected volume of permeate in the tank after a constant interval. The permeate and retentate streams were then reprocessed into the feed tank for a single constituent experiment with betanin.

Experiments were conducted at five different feed concentrations (0.1, 0.2, 0.3, 0.4, 0.5 g L^{-1}), cross-membrane pressures (0.2, 0.3, 0.4, 0.5, 0.6 MPa) and tangential flow rates (200, 300, 400, 500, 600 mL min^{-1}). The range of cross-membrane pressure was selected based on the magnitude of pure water flux and the limitations of the experimental setup. The experimental run for every dataset was performed thrice to check the accuracy of every datum. The betanin concentration in the feed and permeate solutions was determined using a UV spectrophotometer (UV-1800; Shimadzu, Kyoto, Japan) at a wavelength of 534 nm. The standard calibration curve was made using different concentration solutions ranging from 10 to 500 mg L^{-1}.

Initially, the rejection values were checked for two different membranes: HTF-NF 150 and HTF-NF 300, having molecular cut-offs of 150, and 300 Da, respectively (purchased from Permionics Pvt. Ltd). Both membranes showed significantly higher rejection values (>90%). However, the overall flux from the HTF-NF 300 membrane was considerably lower than the HTF-NF 150-membrane. Therefore, the HTF-NF 150 membrane was selected for this particular study. A comparison of the performance of HTF-NF 150 and HTF-NF 300 membranes is reported in the Supporting information (Appendix S1).

Theory

Membrane transport equations

The model given by Spiegler–Kedem can report the equations for the mass transport of permeating species inside the membrane.[23] It is a three-parameter model based on irreversible thermodynamics, which states that all forces acting on permeating species

© 2024 Society of Chemical Industry (SCI).

10974660, 0, Downloaded from https://scijournals.onlinelibrary.wiley.com/doi/10.1002/jctb.7698 by Delhi Technological University, Wiley Online Library on [30/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1. Specification of the test cell

Parameter	Value
Total active membrane area	0.007856 m^2
Membrane diameter	0.1 m
Channel thickness	1 mm
Feed spacer porosity	0.75
Permeate spacer porosity	0.65

are linearly correlated with all fluxes.[24] Equation (1) describes the permeate volumetric flux (J_p) across the membrane.

$$J_p = \frac{L_p}{\rho}(\Delta P - \sigma \Delta \pi) \quad (1)$$

where L_p is the hydraulic permeability, σ is the reflection coefficient, ΔP is the difference in pressure across the membrane and $\Delta\pi$ is the difference in osmotic pressure across the membrane. The linear relationship between osmotic pressure and concentration is given by Eqn (2) as

$$\pi = \alpha c = \frac{\upsilon RT}{M_w} c \quad (2)$$

where α is the constant of proportionality and c is the concentration.

The rejection defines the measure of the separation efficiency of the membrane. The actual rejection (R_j) inside the membrane at a particular point is described in Eqn (3).

$$R_j = \frac{(1-F)\sigma}{1-\sigma F} = 1 - \frac{C_p}{C_m} \quad (3)$$

where C_m and C_p are the concentrations at the membrane's surface on the feed and permeate sides, respectively. Equation (4) gives the flow parameter used in Eqn (3).

$$F = \exp\left(-J_p \frac{1-\sigma}{B_m}\right) \quad (4)$$

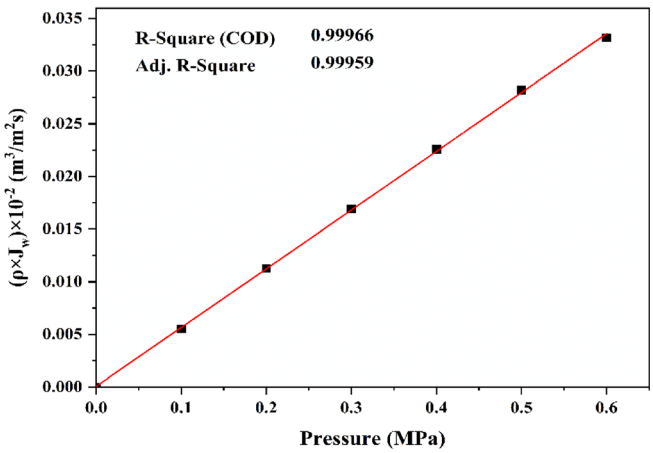


Figure 2. Pure water '$\rho \times J_w$' *versus* ΔP at a constant flow rate of 500 mL min^{-1}.

where B_m is the local solute permeability coefficient of the membrane. Now, the mathematical expression for J_p can be revised in terms of permeate and bulk variables and is given by Eqn (5).

$$J_p = \frac{L_p}{\rho}\left[\Delta P - \sigma \frac{\upsilon RT}{M_w}\phi c_b \left(\frac{(1-F)/\phi}{(1-F)/\phi + (1-\sigma)/\sigma}\right)\right] \quad (5)$$

where ϕ is the concentration polarization and is defined mathematically in Eqn (6).[25-30]

$$\phi = \frac{c_m - c_p}{c_b - c_p} = \exp\left(\frac{J_p}{k}\right) \quad (6)$$

Here, C_b is the bulk solution's solute concentration, and k defines the mass transfer coefficient and is calculated using Eqn (7).

$$k = a \times Q_f^b \quad (7)$$

where Q_f is the volumetric feed flow rate, and a and b are constant coefficients for a given system.

Parameter estimation technique

The above model contains five unknown parameters: the specific hydraulic permeability (L_p), local solute permeability coefficient (B_m), the reflection coefficient (σ) and two mass transfer coefficients (a and b). The actual values of these unknown parameters are required for theoretically evaluating the membrane performance. Values of these parameters were determined by deploying a parameter estimation method and fitting the experimental data in the transport model.[31,32] Appendix S2 represents the general algorithm for the parameter estimation.

Variance-based sensitivity analysis

It is essential to identify the most influential operating and transport parameters to identify the suitability of membranes and optimize the process by controlling the operating parameters. Global sensitivity analysis (Sobol sensitivity analysis) can be used for this purpose, which provides a detailed analysis of which parameter is more sensitive and how the interaction of different input parameters affects the output parameters.[33-35] First-order sensitivity (S) represents the sensitivity of a particular input parameter, and total sensitivity (S_T) represents the sensitivity of that parameter, including the effects of interactions with other parameters.[33-35] Thus, ($S_T - S$) represents the sensitivity effects of interactions of a particular input parameter with other input parameters. The details of the variance-based sensitivity analysis are given in Appendix S3.

RESULTS AND DISCUSSION

All of the experiments and simulations were performed using betanin model solutions with varied feed concentrations at different operating conditions. The cross-membrane pressure was changed from 0.2 to 0.6 MPa, and feed flow rates were changed from 200 to 600 mL min^{-1} at ambient temperature.

Unknown parameter estimation and model validation

Estimation of specific hydraulic permeability (L_p)

The value of L_p does not vary with the solute concentration in the feed.[23,24] Thus, L_p was determined by measuring the pure water flux values at different pressures with a fixed flow rate. Figure 2 represents the graph between the experimental value of pure

© 2024 Society of Chemical Industry (SCI).

10974660, 0, Downloaded from https://scijournals.onlinelibrary.wiley.com/doi/10.1002/jctb.7698 by Delhi Technological University, Wiley Online Library on [30/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2. Comparison of the parameter estimation methods

Method name	Optimization technique	mAPE		MAPE	
		J_P	R%	J_P	R%
Method 1	B_m, σ, a and b optimization using calculated hydraulic permeability	29.49	2.68	11.06	1.05
Method 2	L_p, B_m, σ, a and b optimization	26.40	2.68	7.03	1.02

Abbreviations: J_P, permeate flux; mAPE, maximum absolute percentage error; MAPE, mean absolute percentage error; R%, rejection percentage.

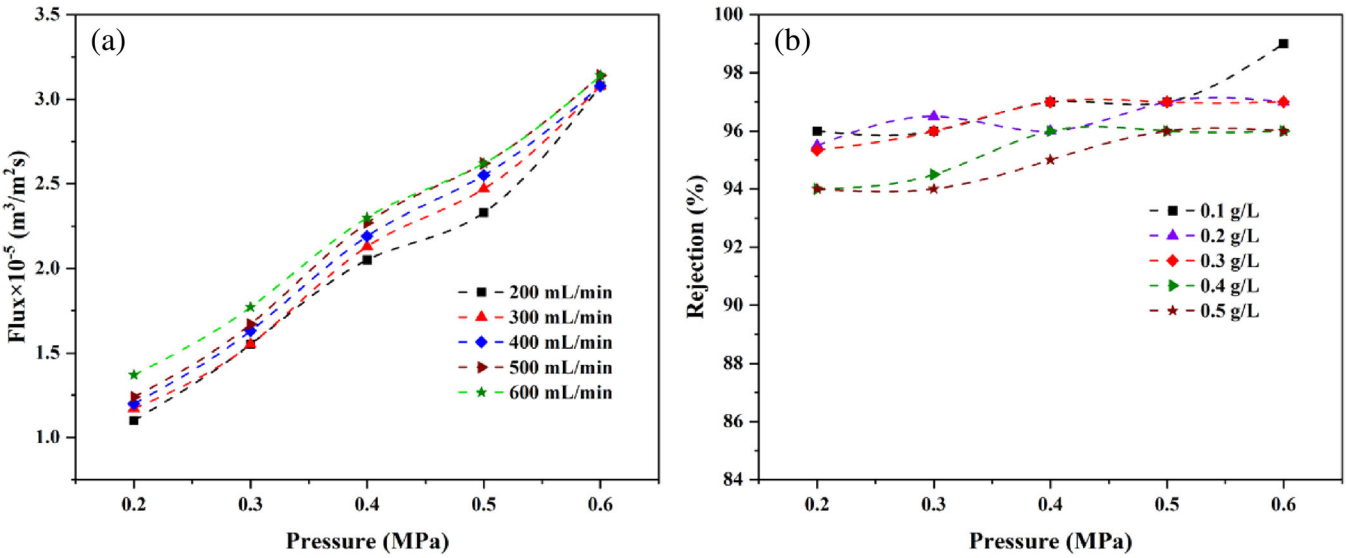


Figure 3. Influence of pressure on (a) permeate flux at a constant feed concentration of 0.1 g L^{-1} and (b) percentage rejection at a constant feed flow rate of 600 mL min^{-1}.

water 'ρ × Jw' *versus* ΔP at a constant flow rate of 500 mL min^{-1}. Experimental results were fitted in the linear equation, and the slope of the curve represents the value of L_p [Eqn (1)]. Experimentally, the value of hydraulic permeability was obtained as 5.52 × 10^{-8} kg m^{-2} s^{-1} Pa^{-1}. The above set of equations now has only four unknown parameters: B_m, σ, a and b.

Estimation of solute permeability (B_m), reflection coefficient (σ) and mass transfer coefficients (a and b)

The values of B_m, σ, and a and b were estimated using different experimental runs at various operating conditions. The cross-membrane pressure was varied from 0.2 to 0.6 MPa, and the feed concentration from 0.1 to 0.5 g L^{-1}. The temperature was maintained at 22 ± 1 °C for the experiment. Feed flow rates of 200, 300, 400, 500 and 600 mL min^{-1} were used for all sets of experimental runs.

The values of different membrane transport parameters (B_m, σ, a and b) were estimated by minimizing the overall error between the theoretical and experimental data. In this experimental work, two optimization methods have been used to analyze the variation of different model parameters on permeate flux and rejection. In the first optimization method, the value of L_p was calculated using the slope of pure water flux *versus* pressure graph. This value was then fixed for optimizing B_m, σ, a and b. The complete data points used for optimization and their model validation data are shown in Appendix S4. For this case, the mean absolute percentage errors for permeate flux and rejection were 11.06% and 1.05%, respectively. All five parameters were optimized in the second optimization method to minimize overall error between theoretical and experimental data points. The mean absolute errors for permeate flux and rejection were 7.03% and 1.02%, respectively. The data for this case is provided in Appendix S5.

Table 2 clearly shows that the optimization of all five parameters (L_p, B_m, σ, a and b) using theoretical and experimental data points (Method 2) showed the lowest mean absolute percentage error for permeate flux and rejection than other optimization methods. Thus, the same values were considered the actual values of these parameters.

Effects of various operating conditions

Various operating conditions, such as feed concentration, cross-membrane pressure and feed flow rates, can influence the performance of the membrane test cell. The roles of these operating conditions were investigated using the same experimental results used for parameter estimation reported in the previous section [Estimation of solute permeability (B_m), reflection coefficient (σ), and mass transfer coefficients (a and b)].

Effect of cross-membrane pressure

Figure 3(a) shows that permeate flux increased with increasing cross-membrane pressure. The results obtained showed that at a constant flow rate of 600 mL min^{-1}, the value of permeate flux

 © 2024 Society of Chemical Industry (SCI).

10974660, 0, Downloaded from https://scijournals.onlinelibrary.wiley.com/doi/10.1002/jctb.7698 by Delhi Technological University, Wiley Online Library on [30/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

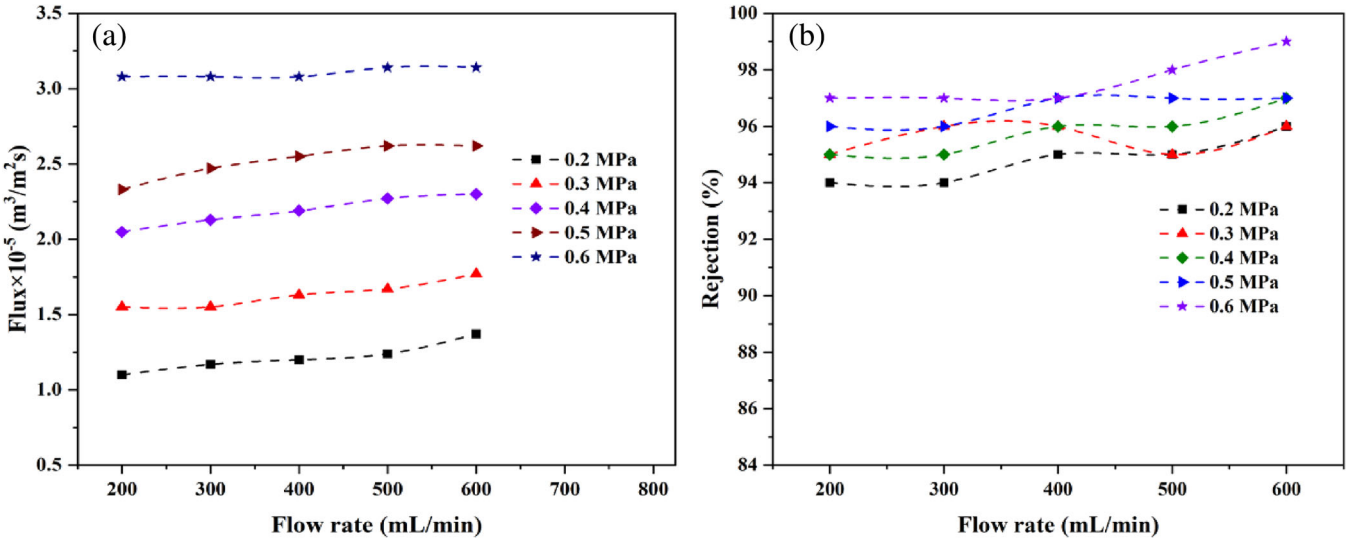


Figure 4. Influence of feed flow rate on (a) permeate flux and (b) percentage rejection, both at a constant feed concentration of 0.1 g L^{-1}.

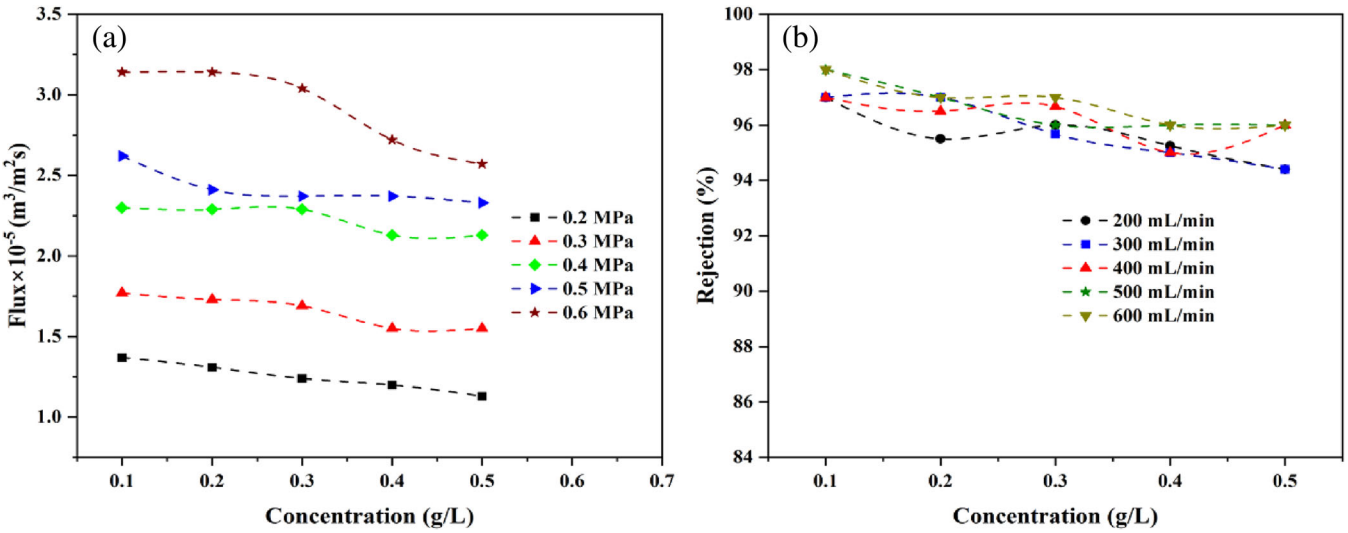


Figure 5. Influence of feed concentration on (a) permeate flux at a constant feed flow rate of 600 mL min^{-1} and (b) percentage rejection at a constant pressure of 0.6 MPa.

increased from 1.37 × 10^{-5} to 3.14 × 10^{-5} m^3 m^{-2} s^{-1} for a feed concentration of 0.1 g L^{-1}, and from 1.13 × 10^{-5} to 2.57 × 10^{-5} m^3 m^{-2} s^{-1} for a feed concentration of 0.5 g L^{-1}. This is likely to have resulted from the increase in the force for permeation, as cross-membrane pressure acts as the main driving force for the mass transfer across the membrane. The rejection percentage increased slightly from 96% to 99% as the cross membrane pressure increased from 0.2 to 0.6 MPa for a feed concentration of 0.1 g L^{-1}, and from 94% to 96% for a 0.5 g L^{-1} feed concentration at a fixed flow rate of 600 mL min^{-1}, as depicted in Fig. 3(b). This may be because an increase in cross-membrane pressure results in the increment of hydraulic flux. However, solute flux does not vary much with varying cross-membrane pressure, as diffusive mass transport dominated over convective mass transport (σ = 0.969), hence resulting in an overall decline in permeate concentration and a slight increase in rejection values.

Effect of feed flow rates

Figure 4(a) depicts the variation of permeate flux as a function of feed flow rates ranging from 200 mL min^{-1} to 600 mL min^{-1} at different pressure variations of 0.2 to 0.6 MPa. It can be seen from the graph that the permeate flux increased from 1.10 × 10^{-5} to 1.37 × 10^{-5} m^3 m^{-2} s^{-1} at a constant pressure of 0.2 MPa and a constant concentration of 0.1 g L^{-1}. The same trend was found for other values of pressure and concentrations. This is because a high feed flow rate reduces the concentration polarization and thus leads to low osmotic pressure differences across the membrane. The values of solute rejection percentage increased from 97% to 99% as the feed flow rate was increased from 200 to 600 mL min^{-1} at a constant pressure of 0.6 MPa and constant concentration of 0.1 g L^{-1} [Fig. 4(b)]. As discussed earlier, a higher flow rate reduces the concentration polarization owing to higher mixing in the feed section, which reduces the osmotic

© 2024 Society of Chemical Industry (SCI).

10974660, 0, Downloaded from https://scijournals.onlinelibrary.wiley.com/doi/10.1002/jctb.7698 by Delhi Technological University, Wiley Online Library on [30/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3. Ranges of different parameters for sensitivity analysis

Parameters	Range
Trans-membrane pressure, MPa	0.2–0.6
Feed concentration, kg m^{-3}	0.1–0.5
Feed flow rate, m^3 s^{-1}	3.3×10^{-6} to 10×10^{-6}
Specific hydraulic permeability coefficient, kg m^{-2} s^{-1} Pa^{-1}	5.5×10^{-9} to 5.5×10^{-7}
Solute permeability, m s^{-1}	4.03×10^{-6} to 4.03×10^{-8}
Reflection coefficient	0.8–1

pressure across the membrane. Osmotic pressure is the driving force for solute transport. Thus, lower osmotic pressure leads to lower solute transport across the membrane and enhances solute rejection. A similar trend was seen for other values of concentration and pressure.

Effects of feed concentration

The values of permeate flux and permeate flow rate reduced with an increase in feed concentration of betanin as higher concentration caused higher osmotic pressure, which decreased the driving force for mass transfer. Figure 5(a) represents the decline in permeate flux value from 3.14×10^{-5} to 2.57×10^{-5} m^3 m^{-2} s^{-1} as the feed concentration increased from 0.1 to 0.5 g L^{-1} at a feed flow rate of 600 mL min^{-1} and a constant pressure of 0.6 MPa. A similar trend was also seen for other values of feed flow rates and pressure. Also, Fig. 5(b) showed the decrease in the rejection percentage from 97% to 94.4% as the feed concentration was increased from 0.1 to 0.5 g L^{-1} at 0.6 MPa pressure and 200 mL min^{-1} feed flow rate. This is a consequence of the lower water and higher solute flux at higher solute concentrations in the feed.

Sensitivity analysis

In order to understand the impact and interdependence of various input variables (C_b, ΔP, Q, L_p, B_m and σ) on output variables (J_p and R_j), the sensitivities of input parameters were analyzed using a suitable range as reported in Table 3. As already discussed in theory (Theory section), variance-based sensitivity analysis requires the formulation of two independent matrices (P and Q) containing equal numbers of input parameter sets. All of the model evaluations were performed in Google Colab. All data were stored in a CSV file for further processing.

The sample size selection was based on the consistency of the first-order and total sensitivity index of various membrane parameters and input variables on output variables. Normalized sensitivities (first-order and total) of different operating conditions and membrane test cell parameters for permeate flux and rejection percentage with variation of sample sizes are represented in SM 6 and SM 7, respectively. As S_i and S_{Ti} were computed after every batch (containing 100 random samples) selected from the whole dataset, convergence occurred slowly after a certain number of samples. The results show that first-order sensitivity and total order sensitivity index fluctuated initially up to 5000 samples, and then varied marginally. By performing different runs, it was found that the sensitivities of all input parameters remained constant after a 50000 sample size, which provides confidence for evaluating the values of sensitivities at a 50000 sample size.

Sensitivity of operating and transport parameters for permeate volumetric flux

Values for S, S_T and their differences ($S_T - S$) were calculated by performing model evaluations in Google Colab (Fig. 6). It was seen that the normalized sensitivities (S and S_T) of specific hydraulic permeability were found to be highest, followed by cross membrane pressure, and sensitivities of all other operating parameters like feed flow rate, feed concentration, solute permeability and reflection coefficient were comparatively negligible. Thus, it can be concluded that hydraulic permeability and cross-membrane pressure are the most influential parameters for total volumetric permeate flux. Results further show that $S_T - S$ is negligible compared to their original total and first-order sensitivity values, indicating that interactions of L_p and ΔP with other input parameters have insignificant effects on the permeate flux. Operating and transport parameters can be sorted in decreasing order as $L_p > \Delta P > C_b > \sigma > B_m > Q$ (Fig. 6(a)).

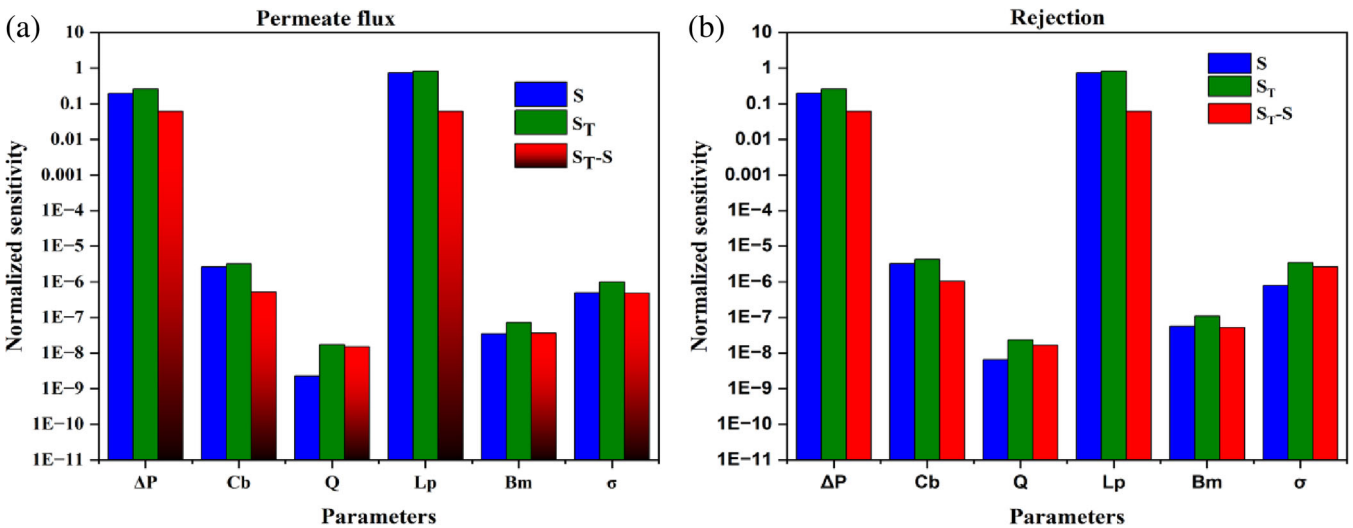


Figure 6. Sensitivity of different operating and membrane module parameters for (a) permeate flux and (b) rejection.

© 2024 Society of Chemical Industry (SCI).

10974660, 0, Downloaded from https://scijournals.onlinelibrary.wiley.com/doi/10.1002/jctb.7698 by Delhi Technological University, Wiley Online Library on [30/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Sensitivity of operating and transport parameters for rejection

Figure 6(b) reports the S, S_T and $S_T - S$ of operating and membrane test cell parameters for rejection. Similar to the case for permeate flux, ΔP and L_p are the most influential parameters. Also, interactions of ΔP and L_p with other transport and operating parameters have negligible effects on the rejection, as shown in Fig. 6(b).

CONCLUSIONS

In this work, the membrane test cell performance of HFT-NF 150 was investigated experimentally and theoretically for the separation of model betanin feed solutions of different concentrations. The predictions for output characteristics obtained theoretically from the Spiegler–Kedem model are well-matched with the experimental results. Low mean absolute percentage error (MAPE) values for output characteristics such as permeate flux and rejection percentage provide confidence for integrating the HFT-NF 150 membrane test cell to design a commercial scale NF unit. Also, the effects of various operating parameters were analyzed in this work. The value of permeate flux increased with the increase in cross-membrane pressure and flow rate but decreased with an increase in concentration. The maximum permeate flux of 3.14×10^{-5} m^3 m^{-2} s^{-1} and a percentage rejection of 99% were obtained at a high flow rate of 600 mL min^{-1} and high pressure of 0.6 MPa, but at a low feed concentration of 0.1 g L^{-1}.

Variance-based sensitivity analysis was also used to analyze the influence and interdependence of operating and membrane test cell parameters on permeate flux and rejection values. The indices began to converge after 5000 and were found to be almost stable at a sample size of 50000. Specific hydraulic permeability and cross-membrane pressure were the most influential parameters for permeate flux and rejection compared to others.

Plant extracts may also contain other bioactive molecules, including proteins, lipids, sugars and minerals, with different molecular sizes. These molecules can also be separated by applying a series of membrane-based processes, such as MF for >100000 Da size particles, UF for micro-size particles, and different cut-off NF membranes for sugars and minerals. A study is required to investigate the separation of these compounds from real plant extracts using membrane-based processes.

Symbols

L_p	Specific hydraulic permeability, kg m^{-2} s^{-1} Pa^{-1}
B_m	Membrane solute permeability coefficient, m s^{-1}
F	Flow parameter
J_p	Permeate flux through membrane, m^3 m^{-2} s^{-1}
c	Solute concentration, kg m^{-3}
k	Mass transfer coefficient, m s^{-1}
a	Coefficient in mass transfer correlation
b	Exponent of Reynolds number in mass transfer equation
d	Membrane test cell diameter, m
P	Pressure, Pa
Q	Flow rate, m^3 s^{-1}
R_j	Rejection
S_c	Schmidt number
S_h	Sherwood number
R_e	Reynolds number
M_w	Molecular weight of solute, kg mol^{-1}
R	Universal gas constant
T	Temperature, K

Greek letters

ϕ	Concentration polarization
υ	Vant Hoff factor
α	Proportionality constant for osmotic pressure
π	Osmotic pressure, Pa
σ	Reflection coefficient
ρ	Density, kg m^{-3}

Subscript

p	Permeate side
b	Bulk side
m	Membrane–feed interface

ACKNOWLEDGEMENT

The authors thank Delhi Technological University for their kind support for the research. No funding was received for this study.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Ashwani Kumar Tiwari: Methodology, Investigation, Data curation, Formal analysis, Visualization, Writing—original draft. Manish Jain: Conceptualization, Writing—review and editing, Supervision, Project administration.

DATA AVAILABILITY STATEMENT

The data supporting this study's findings are available from the corresponding author upon reasonable request.

SUPPORTING INFORMATION

Supporting information may be found in the online version of this article.

REFERENCES

1 Nawaz H, Shi J, Mittal GS and Kakuda Y, Extraction of polyphenols from grape seeds and concentration by ultrafiltration. *Sep Purif Technol* **48**:176–181 (2006). https://doi.org/10.1016/j.seppur.2005.07.006.
2 Wei S, Hossain MM and Saleh ZS, Concentration of rutin model solutions from their mixtures with glucose using ultrafiltration. *Int J Mol Sci* **11**:672–690 (2010). https://doi.org/10.3390/ijms11020672.
3 Pandey KB and Rizvi SI, Plant polyphenols as dietary antioxidants in human health and disease. *Oxid Med Cell Longev* **2**:270–278 (2009). https://doi.org/10.4161/oxim.2.5.9498.
4 Durazzo A, Lucarini M, Souto EB, Cicala C, Caiazzo E, Izzo AA *et al*., Polyphenols: a concise overview of the chemistry, occurrence, and human health. *Phyther Res* **33**:2221–2243 (2019). https://doi.org/10.1002/ptr.6419.
5 Mark R, Lyu X, Lee JJL, Parra-Saldívar R and Chen WN, Sustainable production of natural phenolics for functional food applications. *J Funct Foods* **57**:233–254 (2019). https://doi.org/10.1016/j.jff.2019.04.008.
6 Luna-Guevara ML, Luna-Guevara JJ, Hernández-Carranza P, Ruíz-Espinosa H and Ochoa-Velasco CE, Phenolic compounds: a good choice against chronic degenerative diseases. *Stud Nat Prod Chem* **59**:79–108 (2018). https://doi.org/10.1016/B978-0-444-64179-3.00003-7.
7 Albuquerque BR, Heleno SA, Oliveira MBPP, Barros L and Ferreira ICFR, Phenolic compounds: current industrial applications, limitations, and future challenges. *Food Funct* **12**:14–29 (2021). https://doi.org/10.1039/d0fo02324h.

 © 2024 Society of Chemical Industry (SCI).

10974660, 0, Downloaded from https://scijournals.onlinelibrary.wiley.com/doi/10.1002/jctb.7698 by Delhi Technological University, Wiley Online Library on [30/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

8 Gonalves LCP, Trassi MADS, Lopes NB *et al*., A comparative study of the purification of betanin. *Food Chem* **131**:231–238 (2012). https://doi.org/10.1016/j.foodchem.2011.08.067.

9 Bradley K, Book Review. *Companion Anim* **23**:433 (2018). https://doi.org/10.12968/coan.2018.23.8.433.

10 Alara OR, Abdurahman NH and Ukaegbu CI, Extraction of phenolic compounds: a review. *Curr Res Food Sci* **4**:200–214 (2021). https://doi.org/10.1016/j.crfs.2021.03.011.

11 Zhang QW, Lin LG and Ye WC, Techniques for extraction and isolation of natural products: a comprehensive review. *Chin Med* **13**:1–26 (2018). https://doi.org/10.1186/s13020-018-0177-x.

12 Matta VM, Moretti RH and Cabral LMC, Microfiltration and reverse osmosis for clarification and concentration of acerola juice. *J Food Eng* **61**:477–482 (2004). https://doi.org/10.1016/S0260-8774(03)00154-7.

13 Arboleda Mejia JA, Ricci A, Figueiredo AS, Versari A, Cassano A, Parpinello GP *et al*., Recovery of phenolic compounds from red grape pomace extract through nanofiltration membranes. *Foods* **9**:1–14 (2020). https://doi.org/10.3390/foods9111649.

14 Tonova K, Lazarova M, Dencheva-Zarkova M, Paniovska S, Tsibranska I, Stanoev V *et al*., Separation of glucose, other reducing sugars and phenolics from natural extract by nanofiltration: effect of pressure and cross-flow velocity. *Chem Eng Res Des* **162**:107–116 (2020). https://doi.org/10.1016/j.cherd.2020.07.030.

15 Arend GD, Adorno WT, Rezzadori K, di Luccio M, Chaves VC, Reginatto FH *et al*., Concentration of phenolic compounds from strawberry (Fragaria X ananassa Duch) juice by nanofiltration membrane. *J Food Eng* **201**:36–41 (2017). https://doi.org/10.1016/j.jfoodeng.2017.01.014.

16 Ghosh P, Pradhan RC, Mishra S and Rout PK, Quantification and concentration of Anthocyanidin from Indian blackberry (Jamun) by combination of ultra- and Nano-filtrations. *Food Bioprocess Technol* **11**:2194–2203 (2018). https://doi.org/10.1007/s11947-018-2176-4.

17 Conidi C, Cassano A and Drioli E, Recovery of phenolic compounds from orange press liquor by nanofiltration. *Food Bioprod Process* **90**:867–874 (2012). https://doi.org/10.1016/j.fbp.2012.07.005.

18 Cassano A, Conidi C, Ruby-Figueroa R and Castro-Muñoz R, Nanofiltration and tight ultrafiltration membranes for the recovery of polyphenols from agro-food by-products. *Int J Mol Sci* **19**:1–21 (2018). https://doi.org/10.3390/ijms19020351.

19 Conidi C and Cassano A, Recovery of phenolic compounds from bergamot juice by nanofiltration membranes. *Desalin Water Treat* **56**:3510–3518 (2015). https://doi.org/10.1080/19443994.2014.968219.

20 Mello BCBS, Carlos J, Petrus C and Dupas M, Concentration of flavonoids and phenolic compounds in aqueous and ethanolic propolis extracts through nanofiltration. *J Food Eng* **96**:533–539 (2010). https://doi.org/10.1016/j.jfoodeng.2009.08.040.

21 Arboleda Meija JA, Parpinello GP, Versari A, Conidi C and Cassano A, Microwave-assisted extraction and membrane-based separation of biophenols from red wine lees. *Food Bioprod Process* **117**:74–83 (2019). https://doi.org/10.1016/j.fbp.2019.06.020.

22 Vladisavljević GT, Vukosavljević P and Bukvić B, Permeate flux and fouling resistance in ultrafiltration of depectinized apple juice using ceramic membranes. *J Food Eng* **60**:241–247 (2003). https://doi.org/10.1016/S0260-8774(03)00044-X.

23 Chatterjee A, Ahluwalia A, Senthilmurugan S and Gupta SK, Modeling of a radial flow hollow fiber module and estimation of model parameters using numerical techniques. *J Membr Sci* **236**:1–16 (2004). https://doi.org/10.1016/j.memsci.2004.01.006.

24 Attarde D, Jain M and Gupta SK, Modeling of a forward osmosis and a pressure-retarded osmosis spiral wound module using the Spiegler-Kedem model and experimental validation. *Sep Purif Technol* **164**:182–197 (2016). https://doi.org/10.1016/j.seppur.2016.03.039.

25 Achilli A, Cath TY and Childress AE, Power generation with pressure retarded osmosis : An experimental and theoretical investigation. *J Membr Sci* **343**:42–52 (2009). https://doi.org/10.1016/j.memsci.2009.07.006.

26 Gu B, Kim DY, Kim JH and Yang DR, Mathematical model of flat sheet membrane modules for FO process: Plate-and-frame module and spiral-wound module. *J Membr Sci* **379**:403–415 (2011). https://doi.org/10.1016/j.memsci.2011.06.012.

27 Mccutcheon JR and Elimelech M, Influence of concentrative and dilutive internal concentration polarization on flux behavior in forward osmosis. *J Membr Sci* **284**:237–247 (2006). https://doi.org/10.1016/j.memsci.2006.07.049.

28 Sivertsen E, Holt T, Thelin W and Brekke G, Modelling mass transport in hollow fibre membranes used for pressure retarded osmosis. *J Membr Sci* **417-418**:69–79 (2012). https://doi.org/10.1016/j.memsci.2012.06.014.

29 Xu Y, Peng X, Tang CY, Fu QS and Nie S, Effect of draw solution concentration and operating conditions on forward osmosis and pressure retarded osmosis performance in a spiral wound module. *J Membr Sci* **348**:298–309 (2010). https://doi.org/10.1016/j.memsci.2009.11.013.

30 Gupta VK, Hwang ST, Krantz WB and Greenberg AR, Characterization of nanofiltration and reverse osmosis membrane performance for aqueous salt solutions using irreversible thermodynamics. *Desalination* **208**:1–18 (2007). https://doi.org/10.1016/j.desal.2006.05.023.

31 Attarde D, Jain M, Chaudhary K and Gupta SK, Osmotically driven membrane processes by using a spiral wound module—modeling, experimentation and numerical parameter estimation. *Desalination* **361**:81–94 (2015). https://doi.org/10.1016/j.desal.2015.01.025.

32 Tiwari AK and Jain M, A novel process for rutin recovery from model solutions using nanofiltration: experimental study, mathematical modeling, and scale-up design. *J Food Process Eng* **47**:1–17 (2024). https://doi.org/10.1111/jfpe.14592.

33 Sobol' IM, Sobol 1990_1993. *Mmce* **1**:407–414 (1993).

34 Sharma R and Jain M, Variance-based sensitivity analysis and statistical optimization of design and operating parameters of spiral wound pervaporation modules for thiophene removal from FCC gasoline. *Comput Chem Eng* **141**:106987 (2020). https://doi.org/10.1016/j.compchemeng.2020.106987.

35 Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M and Tarantola S, Variance-based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Comput Phys Commun* **181**:259–270 (2010). https://doi.org/10.1016/j.cpc.2009.09.018.

© 2024 Society of Chemical Industry (SCI).

10974660, 0, Downloaded from https://scijournals.onlinelibrary.wiley.com/doi/10.1002/jctb.7698 by Delhi Technological University, Wiley Online Library on [30/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Communications in Theoretical Physics

ACCEPTED MANUSCRIPT

Constraining Variable Generalized Chaplygin Gas model in Matter Creation Cosmology

To cite this article before publication: Yogesh Bhardwaj *et al* 2024 *Commun. Theor. Phys.* in press https://doi.org/10.1088/1572-9494/ad58c2

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is **© 2024 Institute of Theoretical Physics CAS, Chinese Physical Society and IOP Publishing Printed in China**.

During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.

As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript will be available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by-nc-nd/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 14.139.251.98 on 01/07/2024 at 05:52

Constraining Variable Generalized Chaplygin Gas model in Matter Creation Cosmology

Journal:	*Communications in Theoretical Physics*
Manuscript ID	CTP-240143.R1
Manuscript Type:	Paper
Date Submitted by the Author:	08-May-2024
Complete List of Authors:	Bhardwaj, Yogesh; Delhi Technological University, Applied Mathematics Singh, Prof. Chandra Prakash; Delhi Technological University, Applied Mathematics
Keywords:	Cosmological model, Matter creation, Chaplygin gas model, Dark energy
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
revisedMS.tex	

SCHOLARONE™
Manuscripts

Accepted Manuscript

Constraining Variable Generalized Chaplygin Gas model in Matter Creation Cosmology

Yogesh Bhardwaj & C. P. Singh‡

Department of Applied Mathematics, Delhi Technological University, Bawana Road, Delhi-110042, India.

E-mail: yogeshbhardwaj1597@gmail.com; cpsingh@dce.ac.in

Abstract. We explore the variable generalized Chaplygin gas (VGCG) model in the theory of matter creation cosmology within the framework of a spatially homogeneous and isotropic flat Friedmann - Lemaître - Robertson - Walker space-time. Matter creation cosmology is based on reinterpretation of energy-momentum tensor in Einstein's field equations. This creation corresponds to an irreversible energy flow from the gravitational field to the created matter constituents. The variable Chaplygin gas (VCG) is also studied as a particular solution. We use the Markov Chain Monte Carlo method to constrain the free parameters of three models, namely, ΛCDM, VGCG and VCG models with and without matter creation from the latest observational data from baryon acoustic oscillations, cosmic chronometer, type Ia supernovae (Pantheon) including gamma-ray burst, quasars, and the local measurement of H_0 from R21 data. Two different combinations of dataset provide a fairly tight constraint on the parameters of the ΛCDM, VGCG and VCG models. The present values of various cosmological parameters are obtained which give very close to ΛCDM model. Furthermore, we perform the stability analysis, Bayesian evidence analysis and information criteria analysis for these models through studying the sound speed, Bayes factor, and AIC and BIC selection criterion. The values of sound speed for VGCG and VCG models shows that both the models are stable. According to AIC, it is observed that VGCG and VCG models with matter creation are supported considerably less by current observations whereas BIC shows that these models are not favoured by observational data.

Keywords:Cosmological model, Matter creation, Chaplygin gas model, Dark energy.

Submitted to: *Commun. Theor. Phys.*

1. Introduction

The discovery of the late-time accelerated expansion of the Universe has become one of the most important theoretical challenge in modern cosmology. One of the

‡ Corresponding author

Accepted Manuscript
For Review Only

first observational evidence of this late-time acceleration was shown through the measurement of type Ia supernovae [1–3]. Subsequently, more observations such as Sloan Digital Sky Survey (SDSS) and WMAP [4], baryonic acoustic oscillations (BAO) [5], and cosmic microwave background (CMB) anisotropies [6–10] have confirmed that the Universe is spatially flat and two-thirds of the energy content is dominated by an exotic component with large negative pressure, referred to as dark energy. The simplest candidate for dark energy is the cosmological constant with its equation of state $w = p_\Lambda/\rho_\Lambda = -1$. This model is known as Lambda-Cold dark matter (ΛCDM) model which has three major components: a cosmological constant associated with dark energy, cold dark matter and ordinary matter. This model emerged in the late 1990s as a concordance model. Despite its observational success, ΛCDM model suffers at least from two problems: the cosmological constant problem (CCP) and the cosmic coincidence problem [11, 12]. Many alternative theories have been proposed beyond the ΛCDM model to solve these conundrums. Generally, these alternative theories have been divided into two main classes: First, and possibly the most is the dark energy model in which a new cosmic fluid or matter field is introduced within the framework of general relativity [13]. The other is the modified theories of gravity in which the standard general relativity Lagrangian is modified based on some concrete physical mechanism [14].

We are concerned here with the first kind of dark energy proposal where the dark energy component might be explained by a background fluid with an exotic equation of state, so-called the Chaplygin gas (CG) model [15]. The Chaplygin gas [16] had an origin that was not cosmological. It was used to explain the lifting force on an airplane wing. As the equation of state has negative pressure, it has recently attracted new interest in cosmology. It is used as perfect fluid to describe the accelerating expansion of the Universe with equation of state $p_{Ch} = -A/\rho_{Ch}$, where A is a positive constant, and p_{Ch} and ρ_{Ch} are the energy density and pressure, respectively. The striking feature of CG model is its ability to unify two of the Universe's unknown dark sections: dark matter (DM) and dark energy (DE) [17, 18]. Since the CG model has been ruled out by the observations, this model was extended to the so-called generalized Chaplygin gas (GCG) with the equation of state [19] $p_{Ch} = -A/\rho_{Ch}^\alpha$, where $0 \leq \alpha \leq 1$. This GCG model is also regarded as a unification of DE and DM [20–23]. Gou and Zhang [24] have proposed a variable Chaplygin gas (VCG) model, which explains the evolution from dust to quintessence or phantom. In such model the equation of state is characterized by $p_{Ch} = -A(a)/\rho_{Ch}$, where A is the function of the scale factor a. Yang et al [25], and later on, Lu [26] generalised the VCG model with the equation of state $p_{Ch} = -A(a)/\rho_{Ch}^\alpha$, which is known as variable generalized Chaplygin gas (VGCG) model which also explains the evolution from dust to quintessence or phantom. Some more works on VCG and VGCG models can be found in Refs. [27–32]

Nevertheless, even in the context general relativity, there are other theories to explain the current accelerated expansion of the Universe. As stated above, the presence of negative pressure is the main factor required to accelerate the expansion

Accepted Manuscript

For Review Only

of the Universe. This may also occur, for instance, when the physical system departs from a thermodynamic equilibrium state [33]. In this connection, Zeldovich [34] first stated that the process of cosmological particle creation at the expenses of the gravitational field can be phenomenologically described by a negative pressure. In cosmology, gravitationally induced particle production has a long history. Using the non-equilibrium thermodynamics of open systems, Prigogine et al. [35, 36] proposed the first self-consistent macroscopic formulation of the matter creation process into the Einstein field equations. Later on, Calvâo et al [37] clarified the process through covariant formulation. It has been stated that matter creation can be explained in the realm of the relativistic non-equilibrium thermodynamics. According to several studies [38–47], matter creation cosmology may explain the accelerating expansion of the Universe. The late-time evolution of the Universe has been explained by many authors [48–57]. Since many significant observations have been examined in recent years and the current accelerated expansion of the Universe is satisfactorily explained, the matter creation cosmology is an extremely intriguing topic. The crucial aspect of this model is to take into account the option for the particle production rate, which is still unknown. However, it is assumed to be the function of the Hubble parameter H.

In our previous work [58], we mainly concentrated on the generalized Chaplygin gas inspired by matter creation cosmology and derived the best-fit values of model free parameters. In this paper, to make more complete discussion on contribution of matter creation in CG models, we are interested in discussing the consequences of a variable generalized Chaplygin gas (VGVG) model incorporating the background of matter creation cosmology within the framework of a homogeneous and isotropic flat Friedmann-Lemaître-Robertson-Walker space-time. We also consider a special model of CG, namely, variable Chaplygin gas (VCG). We focus on the analytical solutions and observational analysis of the proposed models to explore the contribution of matter creation. To make our discussion more complete, we use a Bayesian analysis to constrain the parameter space of these models using baryon acoustic oscillations, Cosmic Chronometers data, the Pantheon type Ia supernova including the Gamma Ray Bursts, Quasars and local measurement of H_0 from R21. It is worthy reviewing that the VGCG and VCG models with matter creation can alleviate H_0 tension. We also discuss the stability, Bayesian evidence analysis and selection criterion to distinguish our models with the existing dark energy models such as ΛCDM model.

The paper is organized as follows: In Section 2, for the sake of clarity, we give a brief introduction of matter creation in cosmology. Section 3 is devoted for the background dynamics of the VGCG model in the presence of matter creation. The analytical solutions of VGCG and VCG models are presented in Sections 4 and 5, respectively. Recent observational data and posterior analysis are explored in Section 6. Section 7 is devoted to results and discussion on the consistency of VGCG and VCG models with the observations. Further, Subsections 7.1-7.4 present the evolutions of cosmological parameters, stability, information selection criteria and Bayesian evidence analysis, respectively, and finally, in Section 8, we present the summary of the results.

Accepted Manuscript

For Review Only

2. Basic formalism of matter creation

The thermodynamics of matter creation were first applied to cosmology in the late 1980s by Prigogine and collaborators [35, 36]. They included matter creation explicitly in the matter stress-energy tensor in the Einstein field equations. Later on, Calvão et al [37] discussed and generalized this phenomenological approach through a covariant formulation allowing specific entropy variation. Nowadays, it has been applied in many gravitational theories. In this paper we follow the approach of Prigogine et al [35, 36], which presumes that matter creation takes place at the expense of gravitational energy. The basic formalism of matter creation in theory of general relativity is as follows.

Let us consider an open thermodynamical system where the particle numbers are not conserved. The particle flux vector is assumed to have the form $N^\mu = nu^\mu$, where N is number of particles, n is the particle number density, given by $n = N/V$ and u^μ is the usual particle four velocity. For open systems, the non-conservation of the particle number satisfies the balance equation

$$\nabla_\mu N^\mu \equiv n_{,\mu}u^\mu + 3Hn = n\Gamma, \tag{1}$$

where Γ stands for the rate of change of the particle number in a physical volume $V = V_0a^3$ (where V_0 is the co-moving volume) and $u^\mu_{;\mu} = 3H$ denotes the fluid expansion for FLRW Universe and by notation $\dot{n} = n_{,\mu}u^\mu$. The overdot denotes derivative with respect to the cosmic time. Clearly, $\Gamma > 0$ indicates the creation of particle while $\Gamma < 0$ stands for particle annihilation and $\Gamma = 0$ means there is no creation of matter, i.e., particle number is conserved. The entropy flux vector is given by $s^\mu = n\sigma u^\mu$ [37], where σ is the specific (per particle) entropy. It is to be noted that the entropy must satisfy the second law of thermodynamics $s^\mu_{;\mu} \geq 0$. The incorporation of variable N into the cosmological scenario modifies the first law of thermodynamics as [36, 37]

$$d(\rho V) + pdV - (h/n)d(nV) = 0, \tag{2}$$

where $h = (\rho + p)$ is the enthalpy (per unit volume), and ρ and p are the energy density and thermodynamic pressure, respectively. It is assumed that the Universe evolution continues to be adiabatic, in the sense that the entropy per particle remains unchanged ($\dot{\sigma} = 0$). Equation (2) can be rewritten as

$$d(\rho V) + (p + p_c)dV = 0, \tag{3}$$

where the extra contribution p_c in (3) can be interpreted as a non-thermal pressure, known as creation pressure, defined as

$$p_c = -\frac{h}{n}\frac{d(nV)}{dV}. \tag{4}$$

Using (1) into (4), we get the following form of creation pressure [37]

$$p_c = -\frac{(\rho + p)}{3H}\Gamma. \tag{5}$$

Hence, the creation pressure p_c can be identified with the particle production rate Γ. The simplest possible case, and probably the most physically choice would be one

Accepted Manuscript
For Review Only

for which the characteristic time scale for matter creation is the Hubble time itself. Phenomenologically, this is equivalent to [59, 60]

$$\Gamma = 3\beta H, \tag{6}$$

where β is a positive constant. Using (6) into (5), the creation pressure gives

$$p_c = -\beta(\rho + p). \tag{7}$$

Hereafter, we refer ρ as ρ_{Ch} and p as p_{Ch} for energy density and pressure of Chaplygin gas perfect fluid.

3. Basic field equations

Let us consider the following geometry of the Universe which is very well described by the homogeneous and isotropic flat Friedmann-Lemaître-Robertson-Walker (FLRW) spacetime

$$ds^2 = -dt^2 + a^2(t)\left(dr^2 + r^2 d\theta^2 + r^2 sin^2\theta d\phi^2\right), \tag{8}$$

where $a(t)$ is the scale factor of the Universe. Throughout we use the units such that the speed of light $c = 1$. For comoving observer, $u^\mu = \delta^\mu_\nu$, in which $u^\mu u_\mu = -1$.

The Einstein field equations are written as

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 8\pi G T_{\mu\nu}, \tag{9}$$

where the symbols have their usual meaning. In a general relativistic framework the basic macroscopic variables which describe the thermodynamic states of a relativistic simple fluid are the energy-momentum tensor $T_{\mu\nu}$, the particle flux vector N^μ, and the entropy flux vector s^μ. In this work we shall consider a Universe with matter contents as baryons and Chaplyin gas endowed with matter creation (Chaplygin gas means a perfect fluid having exotic equation of state). Thus, we write the energy density and the pressure of the mixture as a sum of two terms that describe the baryon matter component and Chaplygin gas fluid with matter creation. By taking into account matter creation the effective energy-momentum tensor for the mixture can be given by [37]

$$T_{\mu\nu} = (\rho_{Ch} + \rho_b + p_{Ch} + p_c) u_\mu u_\nu + (p_{Ch} + p_c) g_{\mu\nu}. \tag{10}$$

where ρ_{Ch} and ρ_b are the energy density of Chaplygin gas fluid and baryons, and p_{Ch} and p_c the pressures of Chaplygin gas fluid and particle creation, respectively. The Einstein field equations (9) for line element (8) and energy-momentum tensor (10) yield

$$H^2 = \frac{8\pi G}{3}(\rho_{Ch} + \rho_b), \tag{11}$$

$$\dot{H} + H^2 = -\frac{4\pi G}{3}\left(\rho_{Ch} + \rho_b + 3p_{Ch} + 3p_c\right). \tag{12}$$

where $H = \dot{a}/a$ is the Hubble parameter. In a comoving frame the dot denotes derivative with respect to cosmic time t.

Assuming that baryons are self-conserved, namely the baryon density evolves in

Accepted Manuscript
For Review Only

the normal way, $\rho_b = \rho_{b,0}a^{-3}$, where subscript 0 denotes the value of a quantity at the present time. Thus, the energy densities of the baryon and matter decouple so that the conservation equation $T^{\mu\nu}_{;\nu} = 0$ gives

$$\dot{\rho_{Ch}} + 3H(\rho_{Ch} + p_{Ch} + p_c) = 0. \tag{13}$$

In recent years, it has been suggested that the change of behavior of the missing energy density might be regulated by the change in the equation of state of the background fluid. In this respect Chaplygin gas was introduced as an exotic background fluid with the equation of state described by $p_{Ch} = -A/\rho_{Ch}$, where A is a positive constant [15]. It gives a unified picture of dark matter and dark energy. However, it was ruled out due to inconsistent with the observational data. Subsequently, a more generalized model of the Chaplygin gas is characterized by an equation of state $p_{Ch} = -A/\rho^{\alpha}_{Ch}$ $(0 \leq \alpha < 1)$, which is known as generalized Chaplygin gas (GCG) [17]. It has been shown that this model can be accommodated within the standard structure formation scenarios [17,18]. Therefore, Chaplygin gas models look to be good alternative to explain the accelerated expansion of the Universe.

Guo and Zhang [24] proposed a new Chaplygin gas model, known as variable Chaplygin gas (VCG) model with equation of state $p_{Ch} = -A(a)/\rho_{Ch}$, where $A(a)$ is a function of the scale factor a. In this scenario, the variable Chaplygin gas can drive the Universe from dust to quiescence or phantom. Later on, Yang et al [25] presented a variable generalized Chaplygin gas (VGCG) model whose equation of state is given by [25, 26]

$$p_{Ch} = -\frac{A_0 a^{-n}}{\rho^{\alpha}_{Ch}}, \tag{14}$$

where A_0, n and α are parameters in the model.

In this work, we are interested to study the VGCG model with the cosmology of matter creation along with VCG as particular model.

4. VGCG model with matter creation

Using (7) and (14), Eq. (13) reduces to

$$\rho'_{Ch} + 3(1-\beta)\left(\rho_{Ch} - \frac{A_0 a^{-n}}{\rho^{\alpha}_{Ch}}\right) = 0, \tag{15}$$

where prime denotes derivative with respect to $x = ln(a)$. By taking explicitly the integral of (15), one finds the solution of energy density for VGCG as

$$\rho_{Ch} = \rho_{Ch,0}\left[A_s(1+z)^n + (1-A_s)(1+z)^{3(1-\beta)(1+\alpha)}\right]^{\frac{1}{1+\alpha}}, \tag{16}$$

where $\rho_{Ch,0}$ is the present energy density and $z = (1/a) - 1$ is the redshift and $A_s = \frac{3(1-\beta)(1+\alpha)}{3(1-\beta)(1+\alpha)-n}\frac{A_0}{\rho^{(1+\alpha)}_{Ch,0}}$. For an expanding Universe, we must have $n > 0$ and $(1-\beta)(1+\alpha) > 0$.

Let us investigate the cosmological parameters, specifically the Hubble parameter,

Accepted Manuscript For Review Only

deceleration parameter, equation of state parameter and jerk parameter in order to compare these models with ΛCDM.

Using (16) into (11), the Friedmann equation can also be expressed as

$$H(z) = H_0 E(z), \tag{17}$$

where

$$E(z) = \Big[(1-\Omega_{b,0})\left\{A_s(1+z)^n + (1-A_s)(1+z)^{3(1-\beta)(1+\alpha)}\right\}^{\frac{1}{(1+\alpha)}} + \Omega_{b,0}(1+z)^3\Big]^{1/2}. \tag{18}$$

The acceleration of the Universe is measured by the deceleration parameter which is defined in terms of redshift as

$$q(z) = -1 + (1+z)E(z)^{-1}\frac{dE(z)}{dz}. \tag{19}$$

Using (18) into (19), we obtain

$$q(z) = -1 + \frac{1}{2E^2(z)}\Bigg[3\Omega_{b,0}(1+z)^3 + \frac{(1-\Omega_{b,0})nA_s(1+z)^n + 3(1-A_s)(1-\beta)(1+\alpha)(1+z)^{3(1-\beta)(1+\alpha)}}{\left(A_s(1+z)^n + (1-A_s)(1+z)^{3(1-\beta)(1+\alpha)}\right)^{\frac{\alpha}{1+\alpha}}}\Bigg]. \tag{20}$$

The equation of state (EoS) parameter for the VGCG model is given by

$$w(z) = \frac{-A_s\left(1-\frac{n}{3(1-\beta)(1+\alpha)}\right)(1+z)^n}{A_s(1+z)^n + (1-A_s)(1+z)^{3(1-\beta)(1+\alpha)}}. \tag{21}$$

Lastly, we present the another cosmographic parameter, namely, jerk parameter which is an important tool to distinguish between dynamical models even if they are equivalent kinematically. The degeneracy of the jerk parameter comes from its definition as a third order differential of the scale factor. Its value for the ΛCDM model is 1 (constant), independent of the redshift. In terms of q, it is defined as [8]

$$j(z) = q(2q+1) + (1+z)\frac{dq}{dz}. \tag{22}$$

In forthcoming section, we use the above expression to draw the trajectory of $j(z)$ using best-fit values of parameters and discuss the behavior accordingly.

5. Particular Solution

In this section, we discuss the particular solution of VGCG model. Assuming $n \neq 0$ and $\alpha = 1$ in Eq.(14), the VGCG model reduces to variable Chaplygin gas (VCG) whose EoS is given by $p_{Ch} = \frac{A_0 a^{-n}}{\rho_{Ch}}$. In this model, the energy density is

$$\rho_{Ch} = \rho_{Ch,0}\left[A_s(1+z)^n + (1-A_s)(1+z)^{6(1-\beta)}\right]^{1/2}, \tag{23}$$

Accepted Manuscript

For Review Only

where $A_s = \frac{6(1-\beta)A_0}{\{6(1-\beta)-n\}\rho^2_{Ch,0}}$. Using (23) into the Friedmann equation (11), we obtain

$$H^2(z) = H_0^2\Big[(1-\Omega_{b,0})\left(A_s(1+z)^n + (1-A_s)(1+z)^{6(1-\beta)}\right) + \Omega_{b,0}(1+z)^3\Big]. \quad (24)$$

The deceleration parameter and EoS parameter for this model can be respectively obtained as

$$q(z) = -1 + \frac{1}{2E^2(z)}\Bigg[3\Omega_{b,0}(1+z)^3 + \frac{(1-\Omega_{b,0})nA_s(1+z)^n + 6(1-A_s)(1-\beta)(1+z)^{6(1-\beta)}}{\{A_s(1+z)^n + (1-A_s)(1+z)^{6(1-\beta)}\}^{1/2}}\Bigg], \quad (25)$$

and

$$w(z) = -\frac{A_s\left(1-\frac{n}{6(1-\beta)}\right)(1+z)^n}{A_s(1+z)^n + (1-A_s)(1+z)^{6(1-\beta)}}. \quad (26)$$

6. Observational data

This section addresses the datasets and method to constrain the model parameters of VGCG and VCG models alongwith ΛCDM model. We use the recent observational data of Baryon Acoustic Oscillation, Cosmic Chronometer, Standard candles and Hubble data $R21$ to constrain the model parameters which are described below:

6.1. Baryon Acoustic Oscillations:

We use a set of 17 uncorrelated data points from different Baryon Acoustic Oscillations (BAO) measurements of Refs. [61–74]. The collection of these data points is summarized in Table 1 of Ref. [75]. In the context of BAO, a significant aspect is the computation of the sound horizon at the time of photon-baryon decoupling, r_d, which is defined as

$$r_d = \int_{z_d}^{\infty} \frac{c_s(z)}{H(z)}\, dz, \quad (27)$$

where $c_s(z)$ symbolizes the baryon-photon fluid's sound speed, calculated by $c_s(z) \approx c\left(3 + \frac{9\rho_b(z)}{4\rho_\gamma(z)}\right)^{-0.5}$. Here, $\rho_b(z)$ and $\rho_\gamma(z)$ stand for the densities of baryons and photons, respectively. In observational terms, BAO yields two key projected measures: $\Delta z = \frac{r_d H}{c}$ and $\Delta\theta = \frac{r_d}{(1+z)D_A(z)}$, where Δz is the redshift shift, $\Delta\theta$ is the angular displacement related to BAO and $D_A(z)$ is the angular diameter distance. These measures primarily inform about the product $r_d \times H$. To effectively separate these parameters, independent estimations of the Hubble parameter, specifically H_0, are essential.

Accepted Manuscript For Review Only

6.2. Cosmic Chronometers:

The Cosmic Chronometers (CC) approach provides a method for estimating the Hubble parameter $H(z)$ at different redshifts, leveraging the age difference between passive evolving galaxies. This method is based on the differential age approach, where the relative age of galaxies at different redshifts is related to the inverse of the Hubble parameter. We use a set of 30 uncorrelated CC measurement of $H(z)$, as discussed in Refs. [76–79]. The χ^2 function is defined

$$\chi^2_{CC} = \sum_{i=1}^{30} \frac{[H_{obs}(z_i) - H_{th}(z_i)]^2}{\sigma(z_i)^2}, \tag{28}$$

where $H_{th}(z_i)$ denotes the theoretical value of $H(z)$ at redshift z_i for a given set of cosmological parameters, $H_{obs}(z_i)$ denotes the observed values of $H(z)$ at redshift z_i, and $\sigma(z_i)$ is the uncertainty in the measurement of H_i.

6.3. Standard Candles:

For standard candles (SC), we use uncorrelated measurement of Pantheon type Ia supernovae (SNe) [80], the measurements from Quasars [81] and Gamma Ray Bursts [82]. The Pantheon dataset, follow the luminosity-distance relationship, which can be modeled by $d_L = (1+z)\int_0^z \frac{c}{H(z')}\, dz'$, where d_L is the luminosity distance, z the redshift, c the speed of light, and $H(z')$ is the Hubble parameter as a function of redshift. Gamma-Ray Bursts (GRBs) are utilized by correlating their observed properties, like peak luminosity or total emitted energy, with distance. Their relationship is given by $L_{GRB} = 4\pi d_L^2 F_{obs}$ where L_{GRB} is the intrinsic luminosity and F_{obs} the observed flux. Quasars are considered as the correlation between their luminosity and the size of their broad emission line region, estimated via reverberation mapping. Their distance approximation is $d_L \propto \sqrt{\frac{L_{quasar}}{F_{obs}}}$, with L_{quasar} representing the intrinsic luminosity.

6.4. R21 measurement:

We use $H_0 = 73.04 \pm 1.04\ kms^{-1}Mpc^{-1}$ [83] in our analysis which we denote as $R21$. This is obtained from the Cepheid-SNe sample including systemic uncertainties and lies near the median of all analysis variants which gives a result much closer to the observed in the supernovae sample [84].

We use two different combinations of above datasets. Combination of the first three set of data ($BAO + CC + SC$) is referred as **"BASE"** dataset in our subsequent analysis. The other set of data is termed as **"BASE+R21"**. We use the total likelihood function to derive the joint constraints from the previously mentioned observational data on the model parameter space ($H_0,\ \Omega_b,\ A_s, n,\ \alpha,\ \beta$). In our analysis, we perform a global fitting to determine the cosmological parameters using a Markov Chain Monte Carlo (MCMC). We follow a nested sampler as it is implemented within the open-source package *Polychord* [85] with the *GetDist* package [86] to generate MCMC for

Accepted Manuscript
For Review Only

Table 1. The constraints on the parameters of VGCG and VCG models with matter creation and ΛCDM. The combined dataset of $BAO+CC+SC$ is refereed as "BASE".

Model	Parameters	BASE	$BASE+R21$
ΛCDM	$H_0[kms^{-1}Mpc^{-1}]$	$69.791^{+1.008}_{-1.070}$	$71.432^{+1.111}_{-1.038}$
	Ω_m	$0.271^{+0.018}_{-0.016}$	$0.274^{+0.015}_{-0.018}$
	Ω_Λ	$0.721^{+0.023}_{-0.027}$	$0.726^{+0.022}_{-0.022}$
VGCG	$H_0[kms^{-1}Mpc^{-1}]$	$69.649^{+1.265}_{-2.379}$	$72.674^{+0.355}_{-0.758}$
	Ω_b	$0.021^{+0.012}_{-0.019}$	$0.020^{+0.011}_{-0.019}$
	A_s	$0.758^{+0.027}_{-0.056}$	$0.767^{+0.026}_{-0.046}$
	β	$0.290^{+0.166}_{-0.254}$	$0.260^{+0.155}_{-0.241}$
	n	$0.114^{+0.090}_{-0.110}$	$0.119^{+0.091}_{-0.114}$
	α	$0.509^{+0.332}_{-0.457}$	$0.450^{+0.284}_{-0.392}$
VCG	$H_0[kms^{-1}Mpc^{-1}]$	$69.638^{+1.207}_{-2.491}$	$72.680^{+0.399}_{-0.775}$
	Ω_b	$0.020^{+0.013}_{-0.011}$	$0.020^{+0.013}_{-0.016}$
	A_s	$0.756^{+0.021}_{-0.046}$	$0.765^{+0.021}_{-0.022}$
	β	$0.484^{+0.012}_{-0.025}$	$0.481^{+0.011}_{-0.032}$
	n	$0.105^{+0.088}_{-0.105}$	$0.109^{+0.088}_{-0.105}$

Table 2. The constraints on the parameters of VGCG and VCG models without matter creation. The combined dataset of $BAO+CC+SC$ is refereed as "BASE".

Model	Parameters	BASE	$BASE+R21$
VGCG	$H_0[kms^{-1}Mpc^{-1}]$	$69.674^{+1.217}_{-0.563}$	$72.670^{+1.326}_{-0.412}$
	Ω_b	$0.020^{+0.015}_{-0.024}$	$0.020^{+0.014}_{-0.015}$
	A_s	$0.767^{+0.012}_{-0.033}$	$0.770^{+0.021}_{-0.016}$
	n	$0.129^{+0.101}_{-0.026}$	$0.116^{+0.004}_{-0.011}$
	α	$0.040^{+0.021}_{-0.017}$	$0.042^{+0.072}_{-0.061}$
VCG	$H_0[kms^{-1}Mpc^{-1}]$	$69.870^{+1.133}_{-0.365}$	$72.640^{+1.362}_{-0.352}$
	Ω_b	$0.030^{+0.025}_{-0.036}$	$0.030^{+0.005}_{-0.021}$
	A_s	$0.994^{+0.001}_{-0.002}$	$0.994^{+0.017}_{-0.021}$
	n	$1.140^{+0.032}_{-0.101}$	$0.128^{+0.041}_{-0.035}$

different models using above datasets. We choose the priors for different parameters as $H_0 \in [60;80]$, $\Omega_b \in [0;0.04]$, $A_s \in [0;1]$, $n \in [0;4]$, $\alpha \in [0;1]$, and $\beta \in [0;1]$. We use 100 walkers and 1000 steps to find the best-fit values of parameters. In what follows we perform combined analysis of BAO, CC, SC and R21 on constraints of the ΛCDM, VGCG and VCG models. We use a χ^2 statistic $\chi^2 = \chi^2_{BAO} + \chi^2_{CC} + \chi^2_{SC}$.

Accepted Manuscript
For Review Only

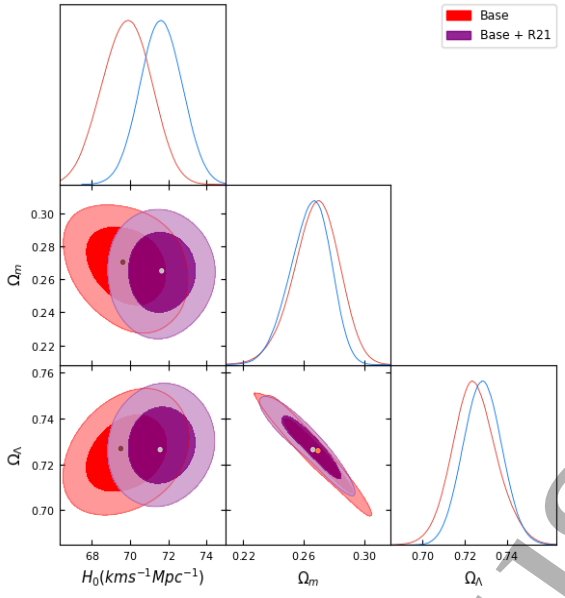


Figure 1. 2D contour at 68.3% and 95.4% confidence levels for ΛCDM model using BASE and $BASE + R21$ data-sets.

Table 3. The present values of different cosmological and geometrical parameters of ΛCDM, VGCG and VCG models with matter creation.

Models→	ΛCDM		$VGCG$		VCG	
Parametrs	BASE	$+R21$	BASE	$+R21$	BASE	$+R21$
z_{tr}	$0.692^{+0.034}_{-0.029}$	$0.689^{+0.028}_{-0.033}$	$0.610^{+0.030}_{-0.022}$	$0.631^{+0.037}_{-0.023}$	$0.674^{+0.029}_{-0.016}$	$0.687^{+0.036}_{-0.031}$
q_0	$-0.591^{+0.01}_{-0.01}$	$-0.590^{+0.02}_{-0.02}$	$-0.541^{+0.02}_{-0.02}$	$-0.542^{+0.05}_{-0.05}$	$-0.553^{+0.03}_{-0.03}$	$-0.555^{+0.03}_{-0.03}$
w_0	$-0.726^{+0.04}_{-0.04}$	$-0.728^{+0.01}_{-0.01}$	$-0.730^{+0.11}_{-0.11}$	$-0.740^{+0.13}_{-0.13}$	$-0.721^{+0.02}_{-0.02}$	$-0.728^{+0.02}_{-0.02}$
$t_0(Gyr)$	$13.73^{+0.02}_{-0.02}$	$13.76^{+0.04}_{-0.04}$	$13.91^{+0.03}_{-0.03}$	$13.94^{+0.07}_{-0.07}$	$13.87^{+0.03}_{-0.03}$	$13.89^{+0.01}_{-0.01}$
j_0	1	1	$0.933^{+0.21}_{-0.21}$	$0.941^{+0.10}_{-0.10}$	$0.979^{+0.02}_{-0.02}$	$0.966^{+0.02}_{-0.02}$
C_s^2	–	–	0.329	0.379	0.730	0.740

Table 4. The present values of different cosmological and geometrical parameters of VGCG and VCG models without matter creation.

Models→	$VGCG$		VCG	
Parametrs	BASE	$+R21$	BASE	$+R21$
z_{tr}	$0.689^{+0.023}_{-0.021}$	$0.694^{+0.019}_{-0.024}$	$0.955^{+0.018}_{-0.023}$	$0.922^{+0.018}_{-0.014}$
q_0	$-0.652^{+0.02}_{-0.02}$	$-0.617^{+0.02}_{-0.02}$	$-0.590^{+0.02}_{-0.02}$	$-0.573^{+0.02}_{-0.02}$
w_0	$-0.589^{+0.11}_{-0.11}$	$-0.595^{+0.15}_{-0.15}$	$-0.838^{+0.01}_{-0.01}$	$-0.842^{+0.02}_{-0.02}$
$t_0(Gyr)$	$13.77^{+0.01}_{-0.01}$	$13.83^{+0.03}_{-0.03}$	$13.91^{+0.02}_{-0.02}$	$13.93^{+0.01}_{-0.01}$
j_0	$0.959^{+0.19}_{-0.19}$	$0.742^{+0.21}_{-0.21}$	$0.892^{+0.11}_{-0.11}$	$1.01^{+0.12}_{-0.12}$
C_s^2	0.414	0.490	0.781	0.805

Accepted Manuscript
For Review Only

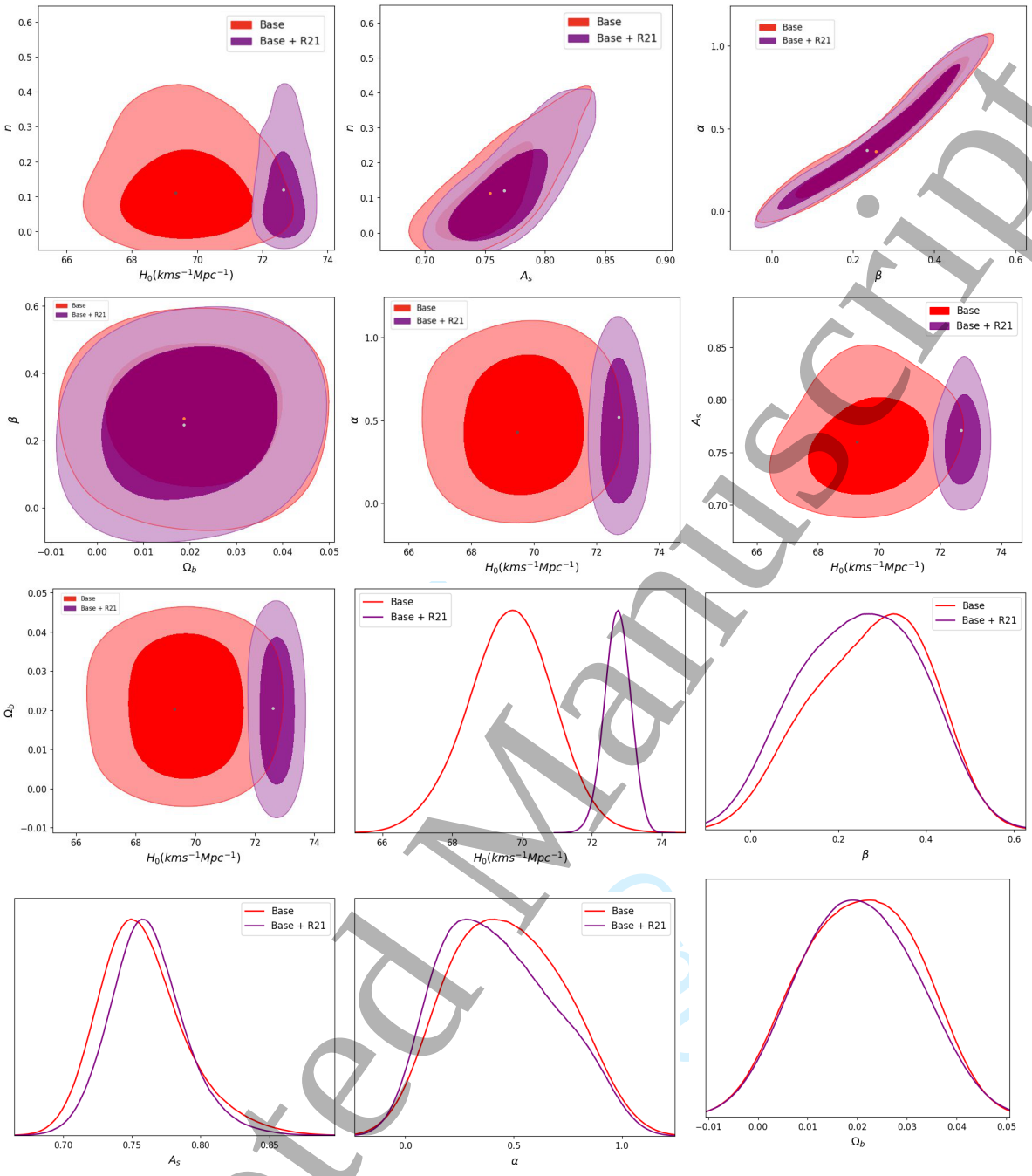


Figure 2. 2D contours at 68.3% and 95.4% confidence levels for pairs (n, H_0), (n, A_s), (α, β), (β, Ω_b), (α, H_0), (A_s, H_0) and (Ω_b, H_0), and 1-D posterior distribution of H_0, β, A_s, α and Ω_b for VGCG model with matter creation using BASE and BASE+R21 datasets, repsectively.

Accepted Manuscript

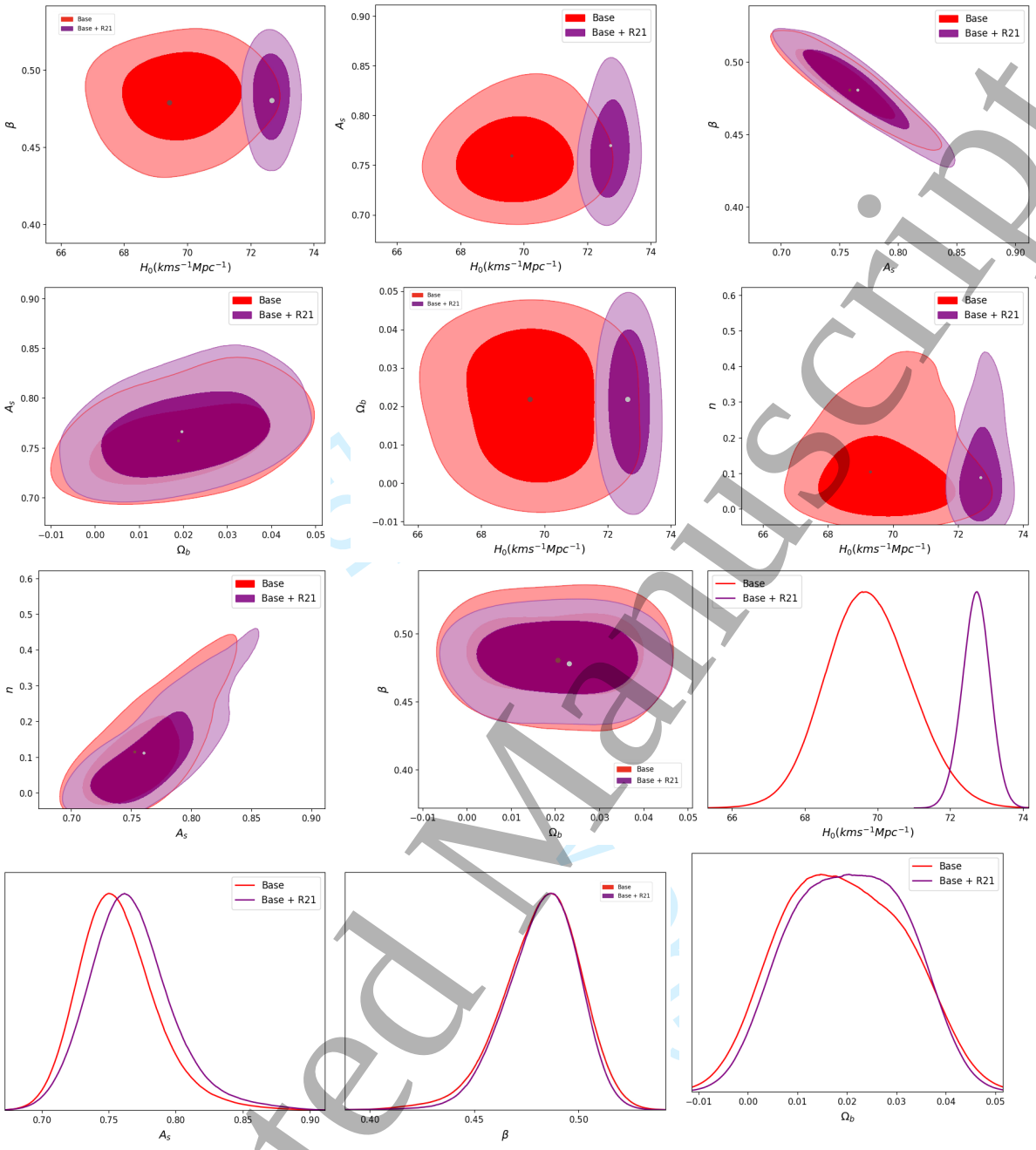


Figure 3. 2D contour at 68.3% and 95.4% confidence levels for pairs (β, H_0), (A_s, H_0), (β, A_s), (A_s, Ω_b), (Ω_b, H_0), (n, H_0), (n, A_s) and (β, Ω_b), and 1-D posterior distribution of H_0, A_s, β and Ω_b for VCG model with matter creation using BASE and BASE+R21 datasets.

Accepted Manuscript

7. Results and discussion

The constraints on different models with and without matter creation for two different combinations of data, namely BASE and $BASE + R21$, are listed in Tables 1 and 2, respectively. As one can see from these tables, the best-fit values of these parameters are nearly same for different dark energy models. It is to be noted that the inclusion of local measurement $H_0 = 73.04 \pm 1.04\ kms^{-1}Mpc^{-1}$ from $R21$ [83] with BASE data, i.e.,$BASE + R21$ always makes the H_0 value higher than BASE data for all the three models discussed above. For ΛCDM, we obtain $H_0 = 71.432^{+1.111}_{-1.038}\ kms^{-1}Mpc^{-1}$ for full datasets, $BASE + R21$ where as for VGCG and VCG, we get $H_0 = 72.674^{+0.355}_{-0.758}$ $kms^{-1}Mpc^{-1}$ and $H_0 = 72.680^{+0.399}_{-0.775}\ kms^{-1}Mpc^{-1}$, respectively. The residual tensions of these fitting results with $R21$ are 0.95σ, 0.36σ and 0.34σ, respectively. On the other hand, we find $H_0 = 69.791^{+1.008}_{-1.070}\ kms^{-1}Mpc^{-1}$, $H_0 = 69.649^{+1.265}_{-2.379}\ kms^{-1}Mpc^{-1}$ and $H_0 = 69.638^{+1.207}_{-2.491}\ kms^{-1}Mpc^{-1}$ with BASE data taken for ΛCDM, VGCG and VCG models, respectively. The $R21$ value of H_0 exhibits a strong tension with the Planck 2018 release $H_0 = 67.4 \pm 0.5$ [10] at 3.92σ CL. Using BASE dataset, we find that the current H_0 tension can be alleviated to 2.19σ in ΛCDM, 1.51σ in VGCG and 1.84σ in the VCG model. In Figures 1-3, we show the constrained contours at 68.3% and 95.4% confidence levels in various parameters plane using two different combinations of data for ΛCDM, VGCG and VGC models, respectively.

In what follows, we discuss the evolutions of different cosmological parameters, stability of the models, selection criteria and Bayesian evidence analysis to compare the proposed models with the existing dark energy models.

7.1. Evolution of cosmological parameters

The evolutions of $H(z)$ with respect to the redshift z for VGCG and VCG with and without matter creation are shown in Figs. 4-7, respectively. The figures also show a comparison of VGCG and VCG models with ΛCDM model. As observed, VCG model closely matches the observed data during the whole evolution whereas the VGCG model coincides during late time evolution of the Universe.

Using BASE and $BASE + R21$ datasets, we obtain the present values of $q(z)$, $w(z)$ and $j(z)$ for ΛCDM, VGCG and VCG models with and without matter creation, respectively which are summarized in Tables 3 and 4. The evolutions of $q(z)$ with redshift z of all the three models for BASE and $BASE + R21$ datasets are shown in Figs. 8 and 9, respectively. We observe that the models show a transition from deceleration ($q(z) > 0$) to acceleration ($q(z) < 0$). In figures, a dot represents the present value of $q(z)$. The present values of $q(z)$ at $z = 0$ from BASE and $BASE + R21$ datasets are $q_0 = -0.591^{+0.01}_{-0.01}$ and $q_0 = -0.590^{+0.02}_{-0.02}$ for ΛCDM, $q_0 = -0.541^{+0.02}_{-0.02}$ and $q_0 = -0.542^{+0.05}_{-0.05}$ for VGCG model and $q_0 = -0.553^{+0.03}_{-0.03}$ and $q_0 = -0.555^{+0.03}_{-0.03}$ for VCG model with matter creation, respectively, as given in Table 3. The values of q_0 show that the models undergo accelerated expansion at present-day. We obtain the transition value z_{tr} where the models transit from deceleration to acceleration at $z_{tr} = 0.610^{+0.030}_{-0.022}$

Accepted Manuscript
For Review Only

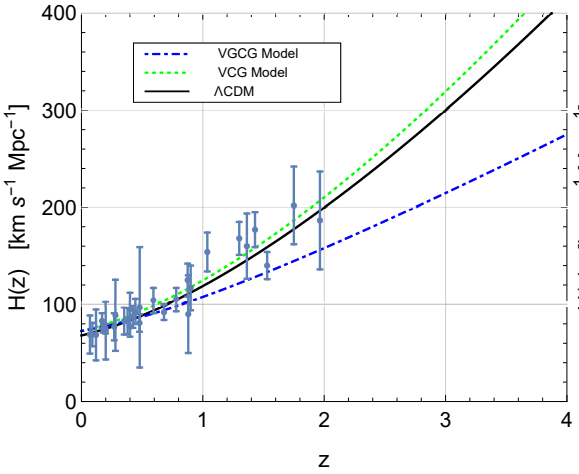


Figure 4. Evolution of $H(z)$ with redshift z in VGCG and VCG models with matter creation, and ΛCDM using 'BASE' data. The grey points with uncertainty bars corresponds to the $H(z)$ sample. The black bold line represents the ΛCDM model.

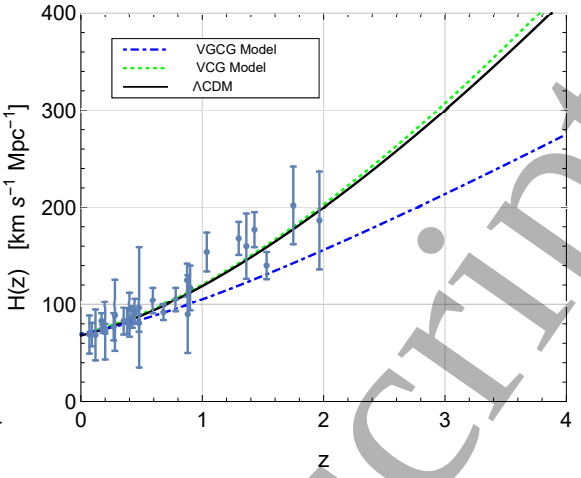


Figure 5. Evolution of $H(z)$ with redshift z in VGCG and VCG models with matter creation, and ΛCDM using 'BASE+R21' dataset. The grey points with uncertainty bars corresponds to the $H(z)$ sample. The black bold line represents the ΛCDM model.

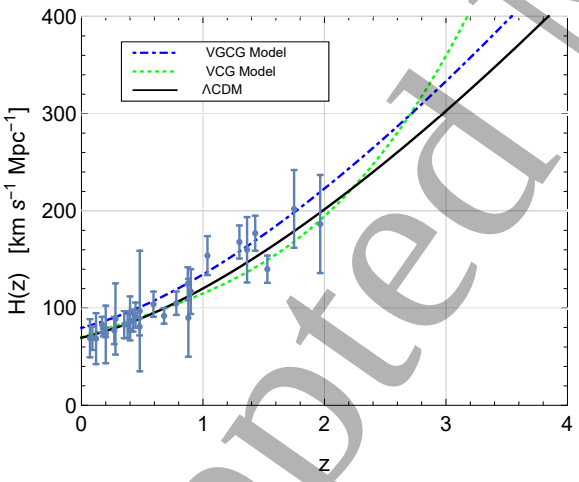


Figure 6. Evolution of $H(z)$ with redshift z in VGCG and VCG models without matter creation, and ΛCDM using 'BASE' dataset. The grey points with uncertainty bars corresponds to the $H(z)$ sample. The black bold line represents the ΛCDM model.

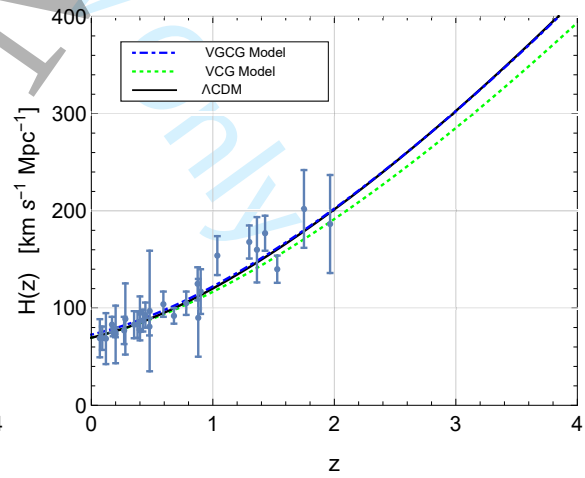


Figure 7. Evolution of $H(z)$ with redshift z in VGCG and VCG models without matter creation, and ΛCDM using 'BASE+R21' dataset. The grey points with uncertainty bars corresponds to the $H(z)$ sample. The black bold line represents the ΛCDM model.

Accepted Manuscript
For Review Only

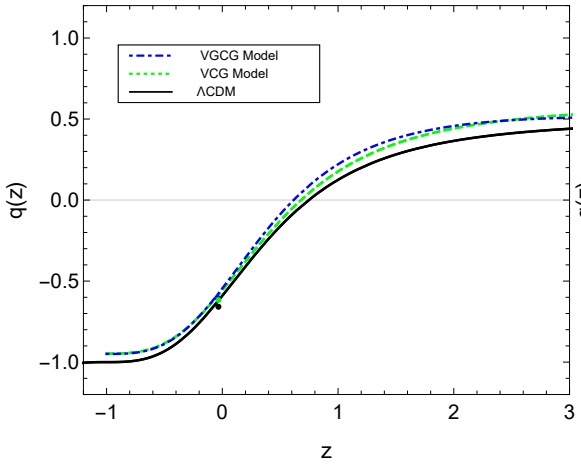


Figure 8. Evolution of $q(z)$ with z in VGCG and VCG models with matter creation, and ΛCDM models using BASE data. The dot line represents the present value of $q(z)$.

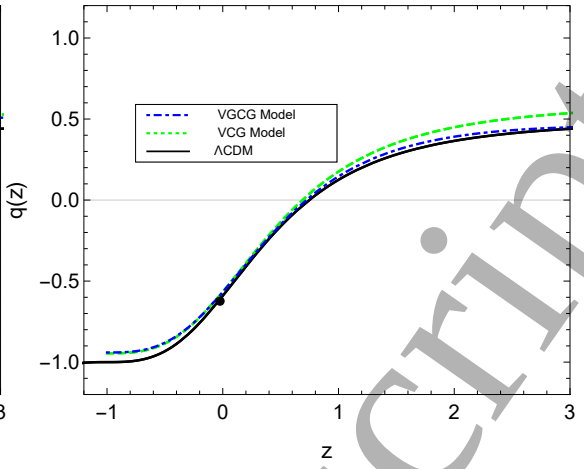


Figure 9. Evolution of $q(z)$ with z in VGCG and VCG models with matter creation, and ΛCDM models using BASE+R21. The dot line represents the present value of q(z).

and $z_{tr} = 0.631^{+0.037}_{-0.023}$ for VGCG model, and $z_{tr} = 0.674^{+0.029}_{-0.016}$ and $z_{tr} = 0.687^{+0.036}_{-0.031}$ in case of VCG model, respectively. These transition values are very close to the ΛCDM transition values (cf. Table 3). We also present the values of z_{tr} and q_0 in Table 4 for these models without matter creation.

The evolutions of EoS parameter $w(z)$ with z of ΛCDM, VGCG and VCG models are shown in Figs. 10 and 11 with two datasets. For both datasets, EoS shows quintessence phase ($-1 < w < 0$) at present and ΛCDM behavior at late epoch. Figures 12 and 13 show the evolutions of the jerk parameter, $j(z)$ with respect to z for the discussed models using BASE and $BASE + R21$ datasets. Based on these two datasets, the present values of $j(z)$ for VGCG model are $j_0 = 0.933^{+0.21}_{-0.21}$ and $j_0 = 0.941^{+0.10}_{-0.10}$ where as for VCG model we get $j_o = 0.979^{+0.02}_{-0.02}$ and $j_0 = 0.966^{+0.02}_{-0.02}$, respectively which are very close to the value $j_0 = 1$ of ΛCDM model.

The ages of Universe for all the models are given in Table 3. The constrains on the age of the Universe for VGCG model are $t_0 = 13.91^{+0.03}_{-0.03}$ Gyr with BASE dataset and $t_0 = 13.94^{+0.07}_{-0.07}$ Gyr with $BASE + R21$ dataset whereas for VCG model we have $t_0 = 13.87^{+0.03}_{-0.03}$ Gyr and $t_0 = 13.89^{+0.01}_{-0.01}$ Gyr, respectively. The ages of the Universe for these Chaplygin gas models are very close to the age of the Universe of ΛCDM model. In Table 4, we list the present values of z_{tr}, q_0, w_0, j_0 and t_0 for the proposed models without matter creation to compare the deviation for constraint results on extended CG models that with matter creation.

Accepted Manuscript
For Review Only

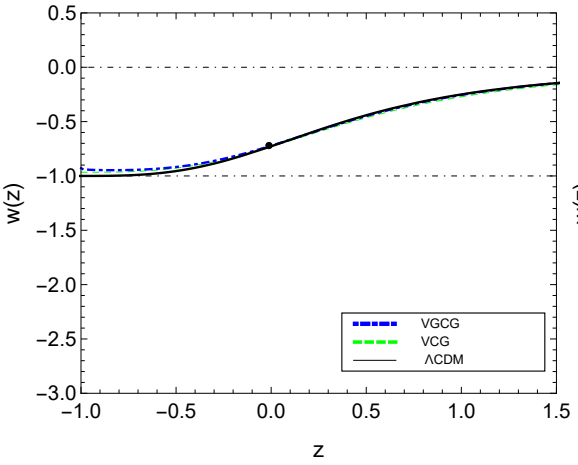


Figure 10. Evolution of $w(z)$ with z in VGCG and VCG models with matter creation, and ΛCDM models using BASE data. The dot line represents the present value of $w(z)$

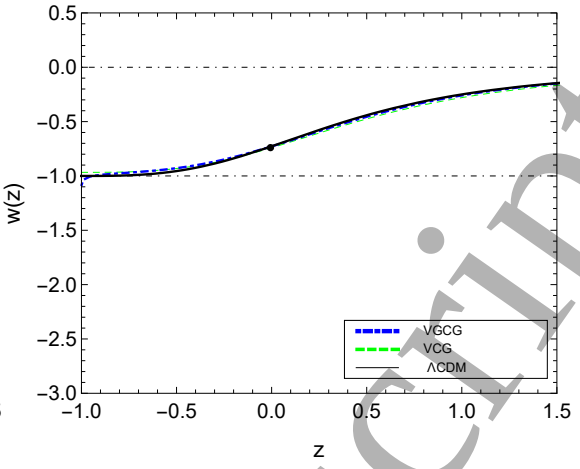


Figure 11. Evolution of $w(z)$ with z in VGCG and VCG models with matter creation, and ΛCDM models using BASE+R21 dataset. The dot line represents the present value of $w(z)$

7.2. Stability of the model

In order to determine the stability of the models, we analyze the squared sound speed of the VGCG and VCG fluids with and without matter creation. For any cosmic fluid model, the squared sound speed $\mathcal{C}_S^2$ can be estimated from the relation: $\mathcal{C}_S^2 = dp_{Ch}/d\rho_{Ch}$. It is noted that $\mathcal{C}_S^2 \geq 0$ corresponds to the stable behavior while $\mathcal{C}_S^2 < 0$ leads to unstable behavior of the model. Thus, the physical values of $\mathcal{C}_s^2$ for stable model must be in the range of (0, 1). Outside of this region, one encounters gradient/tachyonic instabilities and/or instabilities linked to superluminal propagation.

The squared sound speed for VGCG model is given by

$$\mathcal{C}_S^2 = \frac{\alpha A_s(1-\frac{n}{3(1-\beta)(1+\alpha)})(1+z)^n}{A_s(1+z)^n + (1-A_s)(1+z)^{3(1-\beta)(1+\alpha)}}. \tag{29}$$

and for VCG model we have

$$\mathcal{C}_S^2 = \frac{A_s(1-\frac{n}{6(1-\beta)})(1+z)^n}{A_s(1+z)^n + (1-A_s)(1+z)^{6(1-\beta)}}. \tag{30}$$

The sound speed should be positive in the stable model and hence this is justified by taking the constant $\alpha > 0$. Using the best-fit values of model's parameters of BASE and BASE+R21 datasets in above equations, we find the current values $\mathcal{C}_S^2 = 0.329$ and $\mathcal{C}_S^2 = 0.379$ for VGCG model while $\mathcal{C}_S^2 = 0.730$ and $\mathcal{C}_S^2 = 0.740$ for VCG model which are listed in Table 3. We plot the trajectories of $\mathcal{C}_S^2$ with z in Figs.14 and 15 using two combinations of datasets for VGCG and VCG models. The trajectories show that the squared sound speed is always positive through out the evolution lying in $(0, 1)$ and hence positive and less than unity in both the discussed models. This ensure that these

Accepted Manuscript
For Review Only

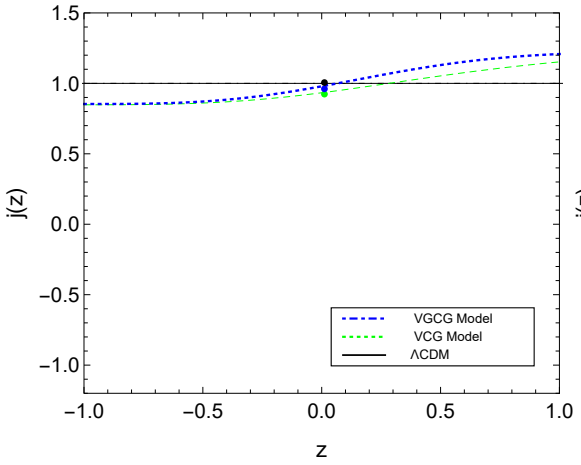


Figure 12. Trajectories of $j(z)$ for VGCG and VCG models with matter creation, and ΛCDM models using best-fit values from BASE data. The dot line represents the present value, j_0

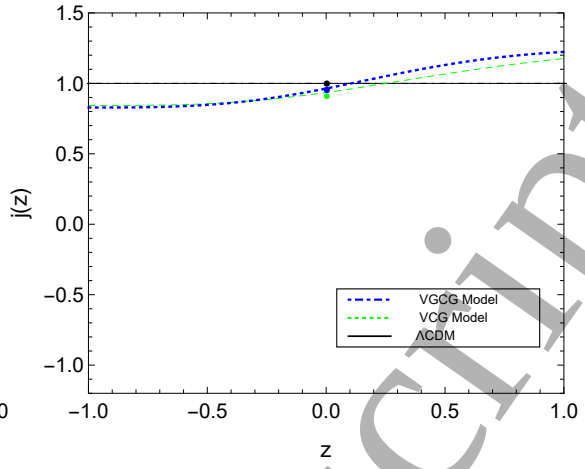


Figure 13. Trajectories of $j(z)$ for VGCG and VCG models with matter creation, and ΛCDM models using BASE+R21 datasets. The dot line represents the present value of j_0.

models do not surpass the velocity of light which confirm their stability. Table 4 shows the $\mathcal{C}_S^2$ for VGCG and VCG models without matter creation. The $\mathcal{C}_S^2$ are positive and less than unity that also confirm the stability.

7.3. Model selection

We calculate the values of χ^2_{min}/dof for ΛCDM, VGCG and VCG models with and without matter creation, which are placed in Tables 5 and 6, respectively. Here, the value of dof (degree of freedom) equals to the number of observational data points (N) minus the number of parameters (K). Here, the number of observational data points for BASE is $N = 273$ (Cosmic Chronometer =30, type Ia supernova(pantheon)=40 and Uncorrelated BAO=17, Hubble diagram for quasars = 24 and Gamma Ray Bursts=162), and $N = 274$ for BASE+R21, and number of parameters are $K = 3, 6$ and 5 for ΛCDM, VGCG and VCG models with matter creation, and $K = 5$ and 4 for VGCG and VCG models without matter creation, respectively. From Tables 5 and 6, we observe that both the models with and without matter creation have $\chi^2_{min}/dof < 1$ for both datasets. This shows that both the models are in good agreement with respect to the observational data as far as χ^2_{min}/dof is concerned.

Although the above analysis are in good agreement with one another, further differences arise due to the choice of cosmological model made.

In what follows, we use the information criteria (IC) which have a deep underpinning in the theory of statistical inference. There are two such IC, namely, Akaike information criteria (AIC) [87] and Bayesian information criteria (BIC) [88], which are respectively

Accepted Manuscript
For Review Only

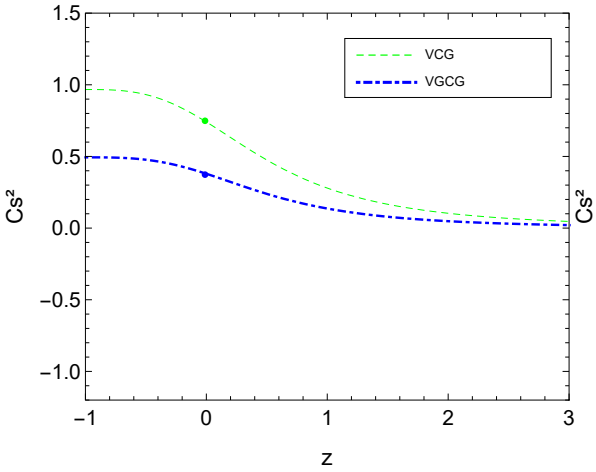


Figure 14. Trajectories of $\mathcal{C}_s^2$ for VGCG and VCG models with matter creation using best-fit values of BASE.

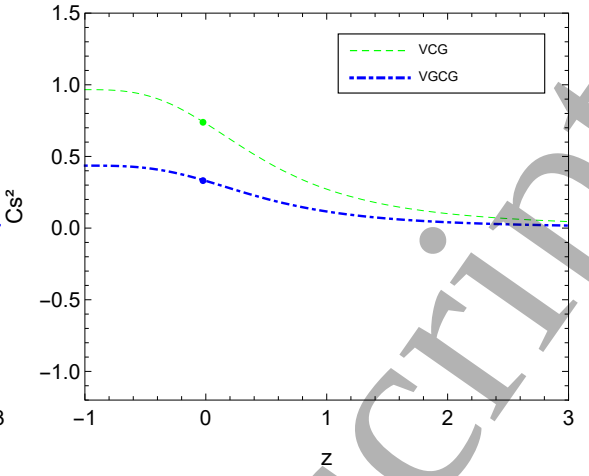


Figure 15. Trajectories of $\mathcal{C}_s^2$ for VGCG and VCG models with matter creation using best-fit values of BASE+R21 dataset.

defined as

$$AIC = \chi^2_{min} + \frac{2NK}{N-K-1}, \tag{31}$$

$$BIC = \chi^2_{min} + K \ln N. \tag{32}$$

where χ^2_{min} measures the quality of the model and K, the number of parameters, interpret model complexity. We calculate the relative values between different models. The one that minimizes the AIC (or BIC) is usually considered the best model. Using ΛCDM as the reference model, we define the AIC and BIC differences as follows: $\Delta AIC = AIC_x - AIC_{\Lambda CDM}$ and $\Delta BIC = BIC_x - BIC_{\Lambda CDM}$, where x represents either VGCG or VCG model. This gives the assessment of strength of the models. The criteria for judging the AIC model [89] are: when $0 \leq \Delta AIC < 2$, the model has almost same support from the data as the best model, for $2 < \Delta AIC \leq 4$, the model is considerably less supported, the range $4 < \Delta AIC < 10$ still supports the considered model. The observational data does not support the model at values of $\Delta AIC > 10$.

In the case of BIC, the range $\Delta BIC < 2$ is considered to support the model from observational data. There is slight evidence against the considered model if $2 < \Delta BIC < 6$. For a value in the range $6 < \Delta BIC < 10$, we obtain strong evidence against the considered model; if $\Delta BIC > 10$, we obtain even stronger evidence against the model [90].

The AIC (BIC) and $\Delta AIC(\Delta BIC)$ for VGCG and VCG models with and without matter creation against the ΛCDM model are listed in Tables 5 and 6, respectively. The ΔAIC values of VGCG and VCG models show that these models are still supported by current observational data. For BIC selection method, it seems that these models are not in favored by observational data. The reason is that AIC is always more general

Accepted Manuscript
For Review Only

Table 5. χ^2_{red}, AIC(BIC), Bayesian evidence and Bayes factor values for ΛCDM, and VGCG and VCG models with matter creation. Here ΛCDM model is taken as a reference model to calculate ΔAIC, ΔBIC and $\ln B$.

Model	Parameters	BASE	$+R21$
ΛCDM	χ^2_{min}	253.92	254.17
	χ^2_{min}/dof	0.94	0.94
	AIC	260.01	260.26
	BIC	270.75	271.00
	$\ln E(M1)$	-123.824	-124.116
VGCG	χ^2_{min}	255.51	257.74
	χ^2_{min}/dof	0.965	0.953
	AIC	267.82	270.05
	BIC	289.16	291.41
	ΔAIC	7.81	9.79
	ΔBIC	18.41	20.41
	$\ln E(M2)$	-125.938	-126.099
	$\ln B_{12}$	2.114	1.983
	Evidence interpret	positive	positive
VCG	χ^2_{min}	254.59	257.63
	χ^2_{min}/dof	0.95	0.96
	AIC	264.81	267.86
	BIC	282.64	285.70
	ΔAIC	4.80	7.60
	ΔBIC	11.89	14.70
	$\ln E(M2)$	-126.513	-126.697
	$\ln B_{12}$	2.689	2.581
	Evidence interpret	moderate	moderate

towards extra parameters whereas BIC tends to penalize them. Hence, the AIC remains useful as it gives an upper limit to the number of parameters.

7.4. Bayesian evidence analysis

In this subsection, we first briefly review the basis of Bayesian inference and model comparison. We then present the computation of the Bayes factor of the two proposed models. The Bayesian evidence applies the same type of likelihood analysis familiar from parameter estimation, but at the level of models rather than parameters. It depends on goodness of fit across the entire model parameter space [91]. It comes from a full implementation of Bayesian inference at the model level, and is the probability of the data given model. It updates the prior model probability to the posterior model

Accepted Manuscript
For Review Only

Table 6. χ^2_{red}, AIC(BIC), Bayesian evidence and Bayes factor values for VGCG and VCG models without matter creation. Here ΛCDM model is taken as a reference model to calculate ΔAIC, ΔBIC and $\ln B$.

Model	Parameters	BASE	$+R21$
VGCG	χ^2_{min}	255.32	257.94
	χ^2_{min}/dof	0.956	0.962
	AIC	265.54	268.16
	BIC	283.36	286.01
	ΔAIC	5.53	7.92
	ΔBIC	12.61	15.01
	$\ln E(M2)$	-126.095	-125.486
	$\ln B_{12}$	2.271	1.370
	Evidence interpret	positive	positive
VCG	χ^2_{min}	254.21	255.76
	χ^2_{min}/dof	0.959	0.950
	AIC	262.36	263.91
	BIC	276.65	278.21
	ΔAIC	2.35	3.67
	ΔBIC	5.90	7.21
	$\ln E(M2)$	-132.818	-132.434
	$\ln B_{12}$	8.994	8.318
	Evidence interpret	decisive	decisive

Table 7. Jeffrey' Interpretive scales for the strength of evidence when comparing two models, $M1$ and $M2$.

$\ln B_{12}$	Strength of evidence
< 1	Inconclusive
$1.0 - 2.5$	Positive evidence
$2.5 - 5.0$	moderate evidence
> 5	decisive

probability using Bayes theorem. The Bayesian evidence is defined as [91–93]

$$E = \int \mathcal{L}(\theta) P(\theta)\, d\theta \tag{33}$$

where θ is the vector of parameters, $\mathcal{L}$ is the likelihood, and $P(\theta)$ is the prior distribution of those parameters before the data were obtained. The evidence of a model is thus the average likelihood of the model in the prior. The Bayesian evidence rewards model predictiveness. Hence the evidence will not penalize the extra parameter in this case, because it does not change the model parameters.

The ratio of the evidence of two models, say M_1 and M_2 is known as the Bayes

Accepted Manuscript
For Review Only

factor which is defined as [94]

$$B_{12} = \frac{E(M_1)}{E(M_2)}, \tag{34}$$

The Bayes factor provides a measure with which to discriminate between the models. It is usual to consider the logarithm of the Bayes factor, for which the so-called Jeffreys scale [95,96] gives empirically calibrated levels of significance for the strength of evidence. Many papers on Bayesian evidence have adopted the interpretive scale suggested by Jeffreys which is presented in Table 7. For each model and dataset we estimate the values of the logarithm of the Bayesian evidence ($\ln E$) and the Bayes factor ($\ln B$) which is the difference between the mean $\ln E$ of ΛCDM model and model concerned. These values are obtained by taking the priors and the dataset described in Sect. 6. We assumed the ΛCDM model as the reference one, say M_1 and therefore its entry is zero by definition.

Table 5 presents the Bayesian evidence and Bayes factor for VGCG and VCG models with matter creation. Looking at the "BASE" and "BASE+R21" datasets respectively, we find $\ln B_{12} = 2.114$ and $\ln B_{12} = 1.983$ for VGCG model, and $\ln B_{12} = 2.689$ and $\ln B_{12} = 2.581$ for VCG model which suggest that VGCG model has positive evidence where as the VCG model has moderate evidence.

Similarly, using the priors and the same datasets, we obtain the Bayes factor for VGCG and VCG models without matter creation which are provided in Table 6. By considering the "BASE" data, we find $\ln B_{12} = 2.271$ for VGCG model and $\ln B_{12} = 8.994$ for VCG model. BASE+R21 data gives $\ln B_{12} = 1.37$ for VGCG model and $\ln B_{12} = 8.318$ for VCG model. It suggests that VGCG model has positive evidence by both the datasets where as the VCG model has strong (decisive)) evidence.

8. Conclusion

In this paper, we have examined the variable generalized Chaplygin gas model inspired by matter creation mechanism in cosmology. Our motivation of this study was in effects of matter creation on the evolution of the Universe filled by variable generalized Chaplygin gas as a perfect fluid using the latest cosmological observations. Specially, we have proposed two Chaplygin gas models with matter creation, i.e., VGCG and VCG and have investigated their evolutionary behaviors via observational datasets. We have placed the constraints on them using two different data combinations: BASE (BAO+CC+SC) and BASE+R21. We have found that the VGCG, VCG and ΛCDM models give a similar χ^2_{min} and correspondingly χ^2_{min}/dof. In what follows, we summarize the results and evolutions of these models.

As stated above, the focus has been on exploring the cosmological models influenced by gravitational matter creation, specifically considering an extended version of the Chaplygin gas with EoS $p_{Ch} = -\frac{A(a)}{\rho^{\alpha}_{Ch}}$ where $0 < \alpha \leq 1$. This approach is utilized to examine the late-stage acceleration of the Universe within the context of a flat FLRW spacetime. Additionally, we have delved into a special scenario where $\alpha = 1$,

Accepted Manuscript
For Review Only

corresponding to the variable Chaplygin gas with EoS $p_{Ch} = -\frac{A(a)}{\rho_{Ch}}$, and incorporated the concept of matter creation. A parallel discussion on these models without matter creation has also been presented.

In the preliminary discussion, it has been explained that matter creation induced by gravity can be regarded as an irreversible scalar phenomenon, contributing an extra decay (or creation) pressure to the energy-momentum tensor in the cosmological fluid mixture. This supplementary creation pressure is characterized by the matter creation rate, denoted as Γ, along with other relevant physical parameters of the fluid. Studies have shown that viewing the Universe as an open system with ongoing matter creation aids in comprehending its current rapid expansion phase. As the nature of the matter creation rate, Γ is still unknown, we have adopted phenomenological approach, assuming Γ evolves as $\Gamma \propto H$. These models were found to be consistently either decelerating or accelerating, lacking a transition from a deceleration phase to an acceleration phase, contradicting the SNe observational data. Nevertheless, our findings indicate that by incorporating the VGCG and VCG equations of state in the matter creation model, a phase transition does occur, even with the aforementioned form of Γ.

The initial section of this manuscript details the analytical solutions derived for matter creation models applying both VGCG and VCG equations of state. Solutions for various cosmological parameters such as energy density, Hubble parameter, deceleration parameter, EoS parameter, and jerk parameter have been deduced as functions of redshift. In the manuscript's latter section, two different joint observational analysis have been performed to constrain the parameters of the model. Tables 1 and 2 present the best-fit values for these model parameters with and without matter creation. We observe that VGCG and VCG models with matter creation can alleviate the current H_0 tension from 3.92σ to 1.51σ and 1.84σ using BASE data. The current values for $q(z)$, $j(z)$, $w(z)$, $t(z)$ and $\mathcal{C}_S^2$ are listed in Tables 3 and 4 and their evolutions have been plotted in Figures 4-13. Our observations indicate a notable variance in the VGCG and VCG models with parameter β from the ΛCDM. These plots vividly illustrate the models' capability to generate late-stage cosmic acceleration while also reflecting a decelerated expansion in earlier times. The transition redshift values, q_0, w_0 and age of the Universe with respect to VGCG and VCG models with and without matter creation are found to be very close to ΛCDM and Planck 2018 release. The evolution trajectories of $j(z)$ reveal minimal deviation from the ΛCDM model for both VGCG and VCG models.

Our analysis employing both AIC and BIC methodologies indicates that models such as VGCG and VCG with and without matter creation exhibit in the range of $4 < \Delta AIC < 10$, suggesting to as much as the preferred model with the joint observational data. In terms of BIC values, the models VGCG and VCG with and without matter creation present $\Delta BIC > 10$, providing evidence against the data used for analysis. It is noteworthy that the squared sound speed for these models lies within the $0 < \mathcal{C}_S^2 < 1$ range, thereby not exceeding the speed of light, which signifies the stability of these models. Consequently, it is deduced that the models discussed in this paper yield cosmological parameters that are well-bounded by the combined observational da-

Accepted Manuscript

For Review Only

ta. Furthermore, these models closely resemble the ΛCDM model in their behavior.

In statistical sense a physical model may be thought as described by a set of parameters. The determination of these parameters may be carried out in many ways; the most commonly used framework to accomplish this is the Bayesian evidence analysis. We have calculated the Bayesian evidence and consequently the Bayes factor to measure the discrimination between the models. Assuming priors for different parameters and ΛCDM model as reference one, we have found that the VGCG model with and without matter creation has positive evidence on Jeffrey scales where as VCG model has moderate evidence with matter creation, and has decisive evidence without matter creation by both datasets.

Acknowledgement

We thank the anonymous referees for useful comments and suggestions that have helped us to improve our manuscript. One of the authors, Y.B. is thankful to Delhi Technological University, New Delhi for providing Research Fellowship to carry out this work.

References

[1] Riess A G et al 1998 *Observational evidence from supernovae for an accelerating universe and a cosmological constant* Astron. J. **116** 1009
[2] Riess A G et al 1999 *BVRI Light Curves for 22 type Ia Supernovae* Astron. J. **117** 707
[3] Perlmutter S et al 1999 *Measurements of Ω and Λ from 42 high redshift supernovae* Astrophys. J. **517** 565
[4] Tegmark M et al 2004 *Cosmological parameters from SDSS and WMAP* Phys. Rev. D **69** 103501
[5] Weinberg D H et al 2013 *Observational probes of cosmic acceleration* Phys. Rep. **530** 87
[6] Bennett C L et al 2013 *Nine - year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and results* Astrophys. J. Suppl. Ser. **208** 20
[7] Ade P A R et al 2014 *Planck 2013 results. XVI. Cosmological parameters* Astron. Astrophys. **571** A16
[8] Blandford R D et al 2004 *Cosmokinetics* ASP Conf. Ser. **339** 27 [astro-ph/0408279]
[9] Ade P A R et al 2016 *Planck 2015 results- XIII. Cosmological parameters* Astron. Astrophys. **594** A13
[10] Aghanim N et al 2020 *Planck 2018 results VI. Cosmological parameters* Astron. Astrophys. **641** A6
[11] Weinberg S 1989 *The cosmological constant problem* Rev. Mod. Phys. **61** 1
[12] Peebles P J E and Ratra B 2003 *The cosmological constant and dark energy* Rev. Mod. Phys. **75** 559
[13] Amendola L et al 2018 *Cosmology and fundamental physics with the Euclid satellite* Living Rev. Rel. **21** 2
[14] Clifton T et al 2012 *Modified gravity and cosmology* Phys. Rept. **513** 1
[15] Kamenshchik A Y, Moschella U and Pasquier V 2001 *An alternative to quintessence* Phys. Lett. B **511** 265
[16] Chaplygin S 1904 *On gas jets* Sci. Mem. Moscow Univ. Math. Phys. **21** 1

Accepted Manuscript
For Review Only

[17] Bento M C, Bertolami O and Sen A A 2002 *Generalized Chaplygin gas, accelerated expansion, and dark energy-matter unification* Phys. Rev. D **66** 043507

[18] Bilić N, Tupper G B and Viollier R D 2002 *Unification of dark matter and dark energy: the inhomogeneous Chaplygin gas* Phys. Lett. B **535** 17

[19] Bento M C, Bertolami O and Sen A A 2004 *Revival of the unified dark energy-dark matter model?* Phys. Rev. D **70** 083519

[20] Bertolami O, Sen A A, Sen S and Silva P T 2004 *Latest supernova data in the framework of the generalized Chaplygin gas model* Mon. Not. R. Astron. Soc. **353** 329

[21] Zhu Z -H 2004 *Generalized Chaplygin gas as a unified scenario of dark matter/energy: Observational constraints* Astron. Astrophys. **423** 421

[22] Santos F C, Bedran M L and Soares V 2006 *On the thermodynamic stability of the generalized Chaplygin gas* Phys. Lett. B **636** 86

[23] Lixin X and Jianbo L 2010 *Cosmological constraints on generalized Chaplygin gas model: Markov Chain Monte Carlo approach* J. Cosmol. Astropart. Phys. **03** 025

[24] Guo Z -K and Zhang Y -Z 2007 *Cosmology with a variable Chaplygin gas* Phys. Lett.B **645** 326

[25] Yang X Y, Wu Y B, LÜ J B and Li S 2007 *Evolution of variable generalized Chaplygin gas* Chin. Phys. Lett. **24** 302

[26] Lu J 2009 *Cosmology with a variable generalized Chaplygin gas* Phys. Lett. B **680** 404

[27] Sethi G et al 2006 *Variable Chaplygin gas: Constraints from CMBR and SNe Ia* Int. J. Mod. Phys. D **15** 1089

[28] Zhang X, Wu F Q and Zhang J 2006 *New generalized Chaplygin gas as a scheme for unification of dark energy and dark matter* J. Cosmol. Astropart. Phys. **01** 003

[29] Wu Ya B, Yang Xiu Yi, Jing H and Fu M H 2008 *The variable generalized Chaplygin gas model with interaction* Mod. Phys. Lett. A **23** 211

[30] Fu M H, Wu Y B and Cheng F Y 2009 *Dynamical stability and attractor of the variable generalized Chaplygin gas model* Chin. Phys. Lett. **26** 069801

[31] Khurshudyan M 2018 *Interaction between generalized varying Chaplygin gas and tachonic fluid* Int. J. Geo. Meth. Mod. Phys. **15** 1850155

[32] Aziza A, Chakraborty G and Chattopadhyay S 2021 *Variable generalized Chaplygin gas in $f(Q)$ gravity and the inflationary cosmology* Int. J. Mod. Phys. D **30** 2150119

[33] Landau L D and Lifshitz E M 1985 *Statistical Physics* (Pergamon, Oxford)

[34] Zeldovich Ya B 1970 *Particle production in cosmology* Pisma Zh. Eksp. Teor. Fiz **12** 443

[35] Prigogine I, Geheniau J, Gunzig E and Nardone P 1988 *Thermodynamics of cosmological matter creation* Proc. Nat. Acad. Sci. **85** 7428

[36] Prigogine I, Geheniau J, Gunzig E and Nardone P 1989 *Thermodynamics and cosmology* Gen. Relativ. Gravit. **21** 767

[37] Calvão M O, Lima J A S and Waga I 1992 *On the thermodynamics of matter creation in cosmology* Phys. Lett. A **162** 223

[38] Steigman G, Santos R C and Lima J A S 2009 *An accelerating cosmology without dark energy* J. Cosmol. Astropart. Phys. **06** 033

[39] Lima J A S, Jesus J F and Oliveira F A 2010 *CDM accelerating cosmology as an alternative to ΛCDM model* J. Cosmol. Astropart. Phys. **11** 027

[40] Basilakos S and Lima J A S 2010 *Constraints on cold dark matter accelerating cosmologies and cluster formation* Phys. Rev. D **82** 023504

[41] Lima J A S, Basilakos S and Costa F E M 2012 *New cosmic accelerating scenario without dark energy* Phys. Rev. D **86** 103534

[42] Fabris J C, de Freitas Pacheco J A and Piattella O F 2014 *Is the continuous matter creation cosmology an alternative to ΛCDM?* J. Cosmol. Astropart. Phys. **06** 038

[43] Lima J A S, Graef L L, Pavón, D and Basilakos S 2014 *Cosmic acceleration without dark energy: background tests and thermodynamic analysis* J. Cosmol. Astropart. Phys. **10** 042

[44] Chakraborty S, Pan S and Saha S 2014 *A third alternative to explain recent observations: Future*

Accepted Manuscript For Review Only

deceleration Phys. Lett. B **738** 424
[45] Rafael C N and Pavón D 2015 *Phantom behavior via cosmological creation of particles* Phys. Rev. D **91** 063526
[46] Harko T, Lobo S N, Mimiso J P and Pavón D 2015 *Gravitational induced particle production through a nonminimal curvature- matter coupling* Eur. Phys. J. C **75** 386
[47] Haro J de and Pan S 2016 *Gravitationally induced adiabatic particle production: From big bang to de Sitter* Class. Quantum Grav. **33** 165007
[48] Pan S, Haro J de, Paliathanasis A and Slagter R J 2016 *Evolution and dynamics of a matter creation model* Mon. Not. R. Astron. Soc. **460** 1445
[49] Saló L A and Haro J de 2017 *Cosmological solutions in spatially curved universes with adiabatic particle production* Class. Quantum Grav. **34** 065001
[50] Pan S, Pal B K and Pramanik S 2018 *Gravitationally influenced particle creation models and late-time cosmic acceleration* Int. J. Geom. Meth. Mod. Phys. **15** 1850042
[51] Singh C P and Kaur S 2019 *Matter creation cosmology in Brans-Dicke theory: Observational tests and thermodynamic analysis* Phys. Rev. D **100** 084057
[52] Singh C P and Kumar A 2020 *Quintessence behavior via matter creation cosmology* Eur. Phys. J. C **80** 106
[53] Singh C P and Kumar A 2020 *Holographic dark energy, matter creation and cosmic acceleration* Phys. Rev. D **102** 123537
[54] Sola J and Yu H 2020 *Particle and entropy production in the running vacuum universe* Gen. Relativ. Gravit. **52** 17
[55] Singh C P and Kaur S 2021 *Constraints on holographic dark energy model with matter creation in Brans-Dicke theory and thermodynamic analysis* Phys. Dark Univ. **33** 100869
[56] Harko T, Lobo F S N and Saridakis E N 2021 *Gravitationally Induced Particle Production through a Nonminimal Torsion-Matter Coupling* Universe **7** 227
[57] Singh C P and Kaur S 2022 *Evolution of holographic dark energy model with adiabatic matter creation* Mod. Phys. Lett. A **37** 2250161
[58] Bhardwaj Y and Singh C P 2024 *Matter creation cosmology with generalized Chaplygin gas* Astrophys. Space Sci. **369** 2
[59] Lima J A S, Germano A S M and Abramo L R W 1996 *FRW-type cosmologies with adiabatic matter creation* Phys. Rev. D **53** 4287
[60] Lima J A S and Alcaniz J S 1999 *Flat Friedmann-Robertson-Walker cosmologies with adiabatic matter creation: Kinematic tests* Astron. Astrophys. **348** 1
[61] Beutler F et al 2011 *The 6dF Galaxy Survey: baryon acoustic oscillations and the local Hubble constant* Mon. Not. R. Astron. Soc. **416** 3017
[62] Ashley J Ross et al 2015 *The clustering of the SDSS DR7 main Galaxy sample-I. A 4 per cent distance measure at z= 0.15* Mon. Not. R. Astron. Soc. **449** 835
[63] Percival Will J et al 2010 *Baryon acoustic oscillations in the Sloan Digital Sky Survey data release 7 galaxy sample* Mon. Not. R. Astron. Soc. **401** 2148
[64] Tojeiro R et al 2014 *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: galaxy clustering measurements in the low-redshift sample of Data Release 11* Mon. Not. R. Astron. Soc. **440** 2222
[65] Blake C et al 2012 *The WiggleZ Dark Energy Survey: joint measurements of the expansion and growth history at* $z < 1$ Mon. Not. R. Astron. Soc. **425** 405
[66] Seo H J et al 2012 *Acoustic scale from the angular power spectra of SDSS-III DR8 photometric luminous galaxies* Astrophys. J. **761** 13
[67] Anderson L et al 2012 *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: baryon acoustic oscillations in the Data Release 9 spectroscopic galaxy sample* Mon. Not. R. Astron. Soc. **427** 3435
[68] Sridhar S et al 2020 *Clustering of LRGs in the DECaLS DR8 footprint: Distance constraints from baryon acoustic oscillations using photometric redshifts* Astrophys. J. **904** 69

Accepted Manuscript
For Review Only

[69] Bautista J E et al 2018 *The SDSS-IV extended baryon oscillation spectroscopic survey: baryon acoustic oscillations at redshift of 0.72 with the DR14 luminous red galaxy sample* Astrophys. J. **863** 110

[70] Abbott T M C et al 2019 *Dark Energy Survey Year 1 results: measurement of the baryon acoustic oscillation scale in the distribution of galaxies to redshift 1* Mon. Not. R. Astron. Soc. **483** 4866

[71] Hou J et al 2021 *The completed SDSS-IV extended baryon oscillation Spectroscopic Survey: BAO and RSD measurements from anisotropic clustering analysis of the quasar sample in configuration space between redshift 0.8 and 2.2* Mon. Not. R. Astron. Soc. **500** 1201

[72] Ata M et al 2018 *The clustering of the SDSS-IV extended Baryon Oscillation Spectroscopic Survey DR14 quasar sample: first measurement of baryon acoustic oscillations between redshift 0.8 and 2.2* Mon. Not. R. Astron. Soc. **473** 4773

[73] Busca N G et al 2013 *Baryon acoustic oscillations in the Lyα forest of BOSS quasars* Astron. Astrophys. **552** A96

[74] Agathe V de S et al 2019 *Baryon acoustic oscillations at $z = 2.34$ from the correlations of Lyα absorption in eBOSS DR14* Astron. Astrophys. **629** A85

[75] Benisty D and Staicova D 2021 *Testing late-time cosmic acceleration with uncorrelated baryon acoustic oscillations dataset* Astron. Astrophys. **647** A38

[76] Moresco M et al 2012 *New constraints on cosmological parameters and neutrino properties using the expansion rate of the Universe to $z \sim 1.75$* J. Cosmol. Astropart. Phys. **07** 053

[77] Moresco M et al 2012 *Improved constraints on the expansion rate of the Universe up to $z \sim 1.1$ from the spectroscopic evolution of cosmic chronometers* J. Cosmol. Astropart. Phys. **08** 006

[78] Moresco M 2015 *Raising the bar: new constraints on the Hubble parameter with cosmic chronometers at $z \sim 2$* Mon. Not. R. Astron. Soc. Lett. **450** L16

[79] Moresco M et al 2016 *A 6% measurement of the Hubble parameter at $z = 0.45$: direct evidence of the epoch of cosmic re-acceleration* J. Cosmol. Astropart. Phys. **05** 014

[80] Scolnic D M et al 2018 *The complete light-curve sample of spectroscopically confirmed SNe Ia from Pan-STARRS1 and cosmological constraints from the combined pantheon sample* Astrophys. J. **859** 101

[81] Roberts C et al 2017 *Tests of Λ CDM and Conformal Gravity using GRB and Quasars as Standard Candles out to $z \sim 8$* arXiv:1711.10369 [astro-ph.CO]

[82] Demianski M et al 2017 *Cosmology with Gamma-ray bursts-I. The Hubble diagram through the calibrated $E_{p,i} - E_{iSO}$ correlation* Astron. Astrophys. **598** A112

[83] Riess A G et al 2022 *A comprehensive measurement of the local value of the Hubble constant with $1 kms^{-1}Mpc^{-1}$ uncertainty from the Hubble Space Telescope and the SH0ES team* Astrophys. J. Lett. **934** L7

[84] Riess A G et al 2019 *Large Magellanic Cloud Cepheid standards provide a 1% foundation for the determination of the Hubble constant and stronger evidence for physics beyond ΛCDM* Astrophys. J. **876** 85

[85] Handley W J, Hobson M P and Lasenby A N 2015 *PolyChord: nested sampling for cosmology* Mon. Not. R. Astron. Soc. Lett. **450** L61

[86] Lewis A 2019 *GetDist: a Python package for analysing Monte Carlo samples* arXiv:1910.13970 [astro-ph.IM]

[87] Akaike H 1974 *A new look at the statistical model identification problem* IEEE Trans. Autom. Control **19** 716

[88] Schwarz G 1978 *Estimating the dimension of a model* Annals Stat. **6** 461

[89] Biesiada M 2007 *Information theoretic model selection applied to supernovae data* J. Cosmol. Astropart. Phys. **02** 003

[90] Liddle A R 2004 *How many cosmological parameters ?* Mon. Not. R. Astron. Soc. **351** L49

[91] Mukherjee P, Parkinson D and Liddle A R 2006 *A Nested Sampling algorithm for cosmological model selection* Astrophys. J. **638** L51

[92] Liddle A R, Mukherjee P, Parkinson D and Wang Y 2006 *Present and future evidence for evolving*

Accepted Manuscript
For Review Only

dark energy Phys. Rev. D **74** 123506
[93] Liddle A R 2007 *Information criteria for astrophysical model selection* Mon. Not. R. Astron. Soc. **377** L74
[94] Efstathiou G 2008 *Limitations of Bayesian evidence applied to cosmology* Mon. Not. R. Astron. Soc. **388** 1314
[95] Jeffreys H 1961 *Theory of probability* 3rd ed. Oxford Classic Texts in the Physical Sciences (Oxford Univ. Press) Oxford
[96] Trotta R 2007 *Applications of Bayesian model selection to cosmological parameters* Mon. Not. R. Soc. **378** 72

Accepted Manuscript
For Review Only

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3405573

Control With Style: Style Embedding-based Variational Autoencoder for Controlled Stylized Caption Generation Framework

Dhruv Sharma, *Chhavi Dhiman, Dinesh Kumar
Department of Electronics and Communication Engineering
Delhi Technological University, Delhi, India
dhruv.0906@yahoo.in, chhavi1990delhi@gmail.com, dineshkumar@dtu.ac.in

***Abstract*— Automatic Image captioning is a computationally intensive and structurally complicated task that describes the contents of an image in the form of a natural language sentence. Methods developed in the recent past focused mainly on the description of factual content in images thereby ignoring the different emotions and styles (romantic, humorous, angry, etc.) associated with the image. To overcome this, few works incorporated style-based caption generation that captures the variability in the generated descriptions. This paper presents a Style Embedding-based Variational Autoencoder for Controlled Stylized Caption Generation Framework (RFCG+SE-VAE-CSCG). It generates controlled text-based stylized descriptions of images. It works in two phases i.e., (*i*) Refined Factual Caption Generation (RFCG), and (*ii*) SE-VAE-CSCG. The former defines an encoder-decoder model for the generation of refined factual captions. Whereas, the latter presents a style embedding-based variational autoencoder for controlled stylized caption generation. The overall proposed framework generates style-based descriptions of images by leveraging Bag-of-Captions (BoC). More so, with the use of a controlled text generation model, the proposed work efficiently learns disentangled representations and generates realistic stylized descriptions of images. Experiments on MSCOCO, Flickr30K, and FlickrStyle10K provide state-of-the-art results for both refined and style-based caption generation, supported with an ablation study.**

***Index Terms*— Bag-of-Captions (BoC), Computer Vision, Controlled Text Generation, Image Captioning, Natural Language Processing, SMU, Stylized Image Captioning, Variational Autoencoder.**

I. INTRODUCTION

IMAGE captioning comprises the description of different objects, their properties, and the interaction between objects in the form of natural language description [1]. With the advent and recent developments in deep-learning techniques, image caption generation has made astounding progress in recent years. Recent methods focus [2, 3, 4, 5] mainly on the factual description of the image, ignoring the inherent sentiments or style within the image. Style-based description of an image provides a way to imitate the language expressing the behavior of human beings. It learns about linguistic styles and utilizes this knowledge to generate style-based image descriptions. Fig. 1, distinguishes between the Factual (F) and style-based (romantic (R), and humorous (H)) caption generation. This example shows that style rich caption provides more reasonable and sentimental descriptions with different opinions or feelings. These artistic captions generated can be further used in numerous applications like storytelling, for visually impaired solutions, visual question answering, etc. and providing better understanding of human expressions. Also, image captions incorporating multiple styles will make visually impaired people perceive images from different perspectives.

Fig. 1: Difference between Factual Image Captions and Stylized Image Captions

In recent past, very few works [6, 7, 8] have explored the generation of the description of images with a specific style. It resulted in the generation of a few publicly available style datasets i.e., Flickrstyle10K [7] , SentiCap [9], and Personality caption dataset [10]. Stylized captioning observes the limitation in generating realistic stylized captions due to limited available paired data i.e. *{Image, factual captions, styled captions-Humorous, Romantic}*. The idea of stylized captioning can be made more generic and robust by developing frameworks independent of the paired set of inputs. To the best of our knowledge, only [10, 11] have attempted to design the frameworks independent of the paired set of inputs. Above mentioned works [6, 7, 8] integrated factual image content with style phrases and rely on paired factual and stylized caption data for training. Ideally, a stylized image captioning model should fulfill two conditions, (*i*) provide the correct description of images, and *(ii)* generates correct stylized words or phrases in correct positions of descriptions. However, the generated sentences do not incorporate sufficient style-related information which is important to describe the content of an image. This may happen due to small size of stylized dataset due to which it is difficult to preserve the correlations between images and captions. This makes the generation of stylized

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 10:10:18 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3405573

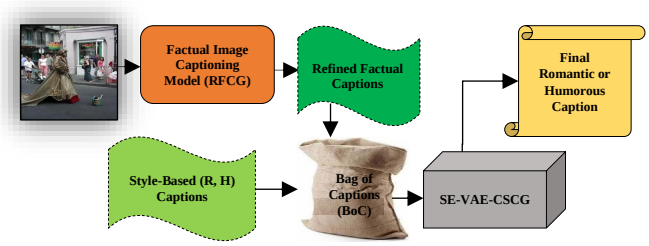


Fig. 2: Block Diagram Representation of the proposed framework

descriptions very difficult. To overcome this challenge and capture the variability in style rich captions, the frameworks should utilize the unpaired set of images and captions, i.e., image is not available with any styled caption. This makes the caption generation acceptable in wider range of applications by not being restricted with the availability of factual and styled captions. In this direction, the proposed work defines Style Embedding-based Variational Autoencoder for Controlled Stylized Caption Generation Framework (SE-VAE-CSCG). It works in two phases. Initially, it generates refined factual captions and forms a Bag-of-Captions by combining both paired and unpaired sets of samples. In the second phase, the BoC feeds the captions to SE-VAE Controlled Caption Generator. Fig. 2 represents a block diagram of the proposed method. The main contributions of this paper are as follows:

(1) An efficient encoder-decoder-based Refined Factual Caption Generation (RFCG) is proposed. It helps the framework to learn from multiple GTs wisely along with visual features, being independent of any personal perspective. RFCG leverages Smooth Maximum Unit (SMU) activated (Squeeze and Excitation Network) SENet at the encoder and soft attention, LSTM, and Bi-LSTM at the language decoder. This improves channel interdependencies at almost no computational cost.

(2) A novel Bag of Captions (BoC) is defined which is a collection of both paired and unpaired samples of caption set ∈ {refined factual, romantic, humorous} for every image sample. It enables the framework to learn and generate stylized captions even if the styles of the image are not highlighted in Ground Truths (GTs).

(3) The proposed SE-VAE-CSCG Framework works in two parts, a modified Variational Autoencoder with Style Embeddings (SE-VAE) and a Deep-LSTM-based generator and discriminator. It learns both unstructured (semantics) and structured (style) feature distributions jointly to generate controlled and plausible stylized captions by updating the style weights.

(4) Extensive experiments on MSCOCO [12], Flickr30K [13], and a subset of FlickrStyle10K [7] demonstrate the efficiency and efficacy of the proposed Refined Factual Image Captioning model and Controlled Style Caption Generation model.

The organization of the rest of the paper is as follows: Section II presents the different state-of-the-art for factual image captioning, stylized image captioning, and controlled text generation. Section III discusses the proposed methodology. Section IV presents the Dataset Used, Implementation details with the results obtained. Finally, Conclusions are drawn in Section V.

II. RELATED WORKS

When This section discusses the related works for factual image captioning, stylized image captioning, and controlled text generation.

A. Factual Image Captioning

Image captioning is a multidisciplinary task that aims to describe the contents of an image in the form of natural language. Traditional methods [14, 15] cannot generate variable-length captions and the descriptions generated from these methods are far from natural. To overcome these limitations deep-learning-based methods [3] [16, 17] are incorporated. Mao et al. [16] proposed a multimodal-RNN structure that utilized deep-CNN networks with m-RNN to generate novel sentence descriptions. Vinyals et al. [17] proposed an end-to-end framework that generated qualitative and quantitative results by generating sentences in plain English. Further, an extension to LSTM is presented in [18] coined as g-LSTM. This model inputted extra semantic information from the image and generated variable-length natural sentences for a particular image. Inspired by the recent development in attention mechanisms in deep learning, various models adopted different attention mechanisms [2, 3] etc. Further, few models [19, 20] paid more attention to salient objects for the generation of corresponding words. Lu et al. [21] proposed an adaptive attention-based captioning model with visual sentinel. This model decides whether to pay attention to the image or the visual sentinel. The work in [22] presented a modified beam search strategy for concept-augmented image captioning with fused YOLO9000 and ResNet features.

With the advancements in deep-learning, the transformer-based models [5] [23] are being widely used for image captioning task. These models captured relative spatial relationship between image regions and provided a better computational complexity. Also, Li et al. [24] proposed entangled transformer to extract more discriminative features by considering novel objects. Pan et al. [25] proposed X-linear attention network for image captioning that leveraged 2^{nd} and higher order interactions by exploiting low-rank bilinear pooling. In recent times, Large Language Models (LLMs) demonstrated outstanding zero and few-shot performance on a range of tasks like translation, text summarization, and common-sense reasoning. These models are able to learn the complex relationships between visual characteristics and language expressions because they have been trained on large datasets that contain corresponded with images and captions. Researchers utilized the concept of LLM in image captioning tasks to improve caption generation and to get enriched captions [26, 27]. LLM-based captioning models still require more finer visual details as the caption generated does not offer more detail. Also, the LLM-based captioning models sometimes fails

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 10:10:18 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3405573

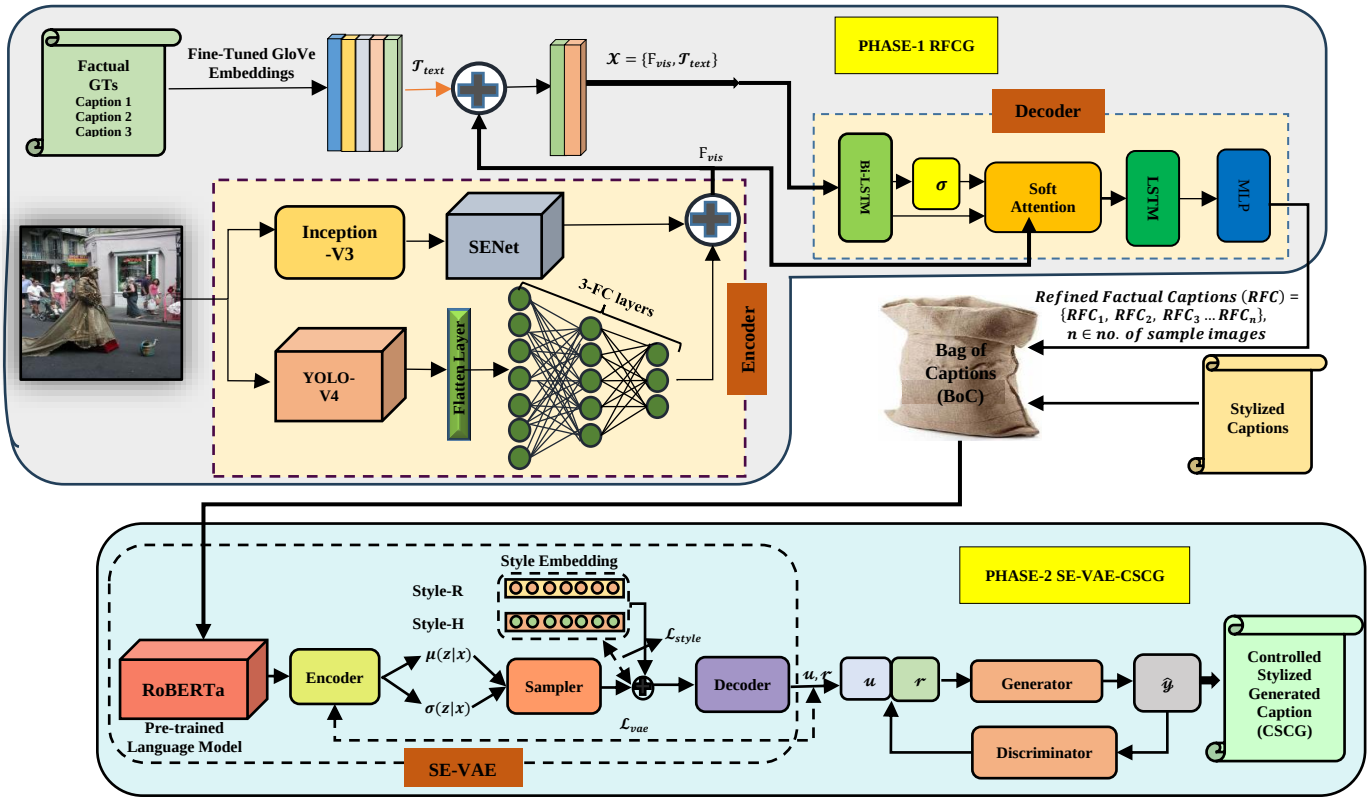


Fig. 3: Illustration of the proposed Control with style Framework: Phase-I RFCG Module, and Phase-II SE-VAE-CSCG Module. In Phase-I, RFCG module encoder generates visual embeddings as F_{vis} for the input image. Whereas RFCG module decoder receives combination of text embeddings $\mathcal{T}_{text}$ and the visual embeddings F_{vis} given by $\mathcal{X} = \{F_{vis}, \mathcal{T}_{text}\}$, generating refined captions as Refined Factual Captions (RFC) = $\{RFC_1, RFC_2, RFC_3 \ldots RFC_n\}$, n ∈ no. of sample images. From Phase I, a Bag of Captions (BoC) is defined that leverages Style Embedding based Variational Auto-Encoder-CSCG (SE-VAE-CSCG) in Phase-2 for generation of controlled stylized captions.

to elicit a judgment or rarely, generate malformed image descriptions. This may omit semantically important image details as LLM-based fusion process sometimes ignores inter-element dependencies. To overcome this issue in captioning model, this paper fuses the factual knowledge from images with the stylized knowledge to generate coherent and diverse image captions.

B. Stylized Image Captioning

Stylized image captioning aims to generate captions that incorporate different styles or sentiments in the captions. This technique focuses not only on the factual content but also on the linguistic style content in the image. Gan et al. [7] proposed a novel framework that addressed the generation of captions with styles for images and videos. This model incorporated style factored-LSTM modules and multi-task learning. Also, Chen et al. [6] presented a style-factual LSTM module that captured factual and style knowledge for the generation of style-based descriptions of images. Another work proposed an adversarial learning model named MSCap [28] for multi-style image captioning. Existing methods are based on styles trained captions aligned to images, generating captions with low relevance. To overcome this limitation Mathews et al. [29] proposed SemStyle to generate visually relevant styled captions from a large corpus of styled text without paired images. Sentimental Image Captioning [8] automatically incorporated sentiments within images using attention LSTM. Further, fluent and relevant descriptions are generated with inherent sentimental styles and language LSTM. Li et al. [11] proposed a novel data augmentation framework that extracted style phrases from small-scale stylized sentences and provides factual descriptions from them. Another work, [30] proposed two-step transformer-based method for style-based caption generation: first, separate style presentations from substantial corpora of stylized text; then, join the separated style representations to the image content. Wu et al. [31] generated stylized sentences in a multi-pass decoding process using reinforcement learning paradigm. Few researchers [32, 33] also focused on LLM model based stylized caption generation. Though these methods have shown promising results, they still require sufficient samples with desired style labels, and additional training. However, these methods are less flexible as they require paired prompts that represent the relations and differences between different styles. To address this concern, the paper uses paired style and factual descriptions for an image to multiple stylized descriptions with the use of unpaired data that also captures variability in the generated captions. With this the proposed model also provided an improvement in style accuracy as the proposed framework incorporates text-only data for stylized captioning. This enables the proposed framework to learn task-specific details and subtleties that are challenging for fully zero-shot methods.

C. Controlled Text Generation

Deep generative models have gained immense popularity in recent years. Hu et al. [34] proposed a formal connection

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 10:10:18 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3405573

between GANs and VAEs which provided a powerful tool to analyze a diverse set of existing model variants. Goodfellow et al. [35] introduced GAN that simultaneously learns the generator-discriminator model and provided improvements in the performance of the classification models for limited labeled data. Larsen et al. [36] combined GAN and VAE to learn feature representation in the discriminator module based on the VAE reconstruction objective for enhanced image generation. Due to the non-differentiability of discrete samples in text generation, discriminator-based modules are hard to implement. Few works [37] [38] focused on VAEs without discriminators for the controlled generation of text. These models provided good results when compared with LSTM language models but still, they lack disentangled latent representations. To overcome this limitation, Chen et al. [39] presented a GAN model that learned disentangled representation in an unsupervised manner. Further, Siddharth et al. [40] discussed VAEs-based semi-supervised learning for disentangled representation of images. In another work, Hu et al. [41] presented a controllable text generation model which generated plausible text sentences controlled by learning disentangled latent representations with designated semantics. Intending to generate sentences with semantic coherence Betti et al. [42] presented CTERM-GAN which assesses the coherence with the external conditioning. To provide improvements in interpretability this paper focused on the disentangled representation of text with consistent style or sentiments.

III. PROPOSED METHODOLOGY

This section discusses a proposed framework which is sub-divided into two sub-sections namely: (1) Refined Factual Caption Generation (RFCG), and (2) Stylized Image Captioning using SE-VAE, and modified controlled text generation module. The two-phase proposed framework is depicted in Fig. 3. To generate a style-based description of images first the factual descriptions are extracted and generated factual captions are fed to the SE-VAE-CSCG. This model is able to generate stylized descriptions with unpaired style transfer i.e., for a given image there is no need for corresponding factual and stylized captions. Also, this model does not require any image features for the generation of style-based descriptions. The detailed structure of the two models is discussed below:

A. Refined Factual Caption Generation

Refined Factual Image Caption Generation (RFCG) utilizes an encoder-decoder structure to generate refined factual captions of images using the available factual GTs per image. It helps the model to produce human-independent (unbiased) refined captions by merging both visual features (global features and object-level local features) with finetuned GloVe embeddings-based text encodings in the encoder. The decoder learns the combined visual and textual encodings to generate the most likely refined factual descriptions unbiased from human perception using Bi-LSTM and LSTM layers.

(i) RFCG- Encoder

In RFCG-Encoder, Inception-V3 [43] extracts global spatial features of images. Symmetric and asymmetric modules in 42 layered architecture of Inception V3 [43] provide low error rates with high efficiency when compared with its previous versions and its contemporaries. A feature vector of dimensions $(10 \times 10 \times 2048)$ extracted from the Inception-V3 is passed to the Smooth Maximum Unit (SMU) activated Squeeze and Excitation Network (SENet) [44] to improve the channel interdependencies with no added computational cost. It performs channel scaling with learned weighted features as f_{vec}. We have used SMU [45] activated features to map with original channels, that has the potential to provide performance improvements when compared with traditional activation functions i.e., ReLU. SMU can be realized using the smooth activation of the maximum function which can smoothly approximate the ReLU activation. The internal architecture is provided in Fig. 4. Mathematically, squeeze and excitation operations are represented by:

$$z_{se} = \mathbb{f}_{sq} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} m_{se}(i,j) \quad (1)$$

$$s = \mathbb{f}_{ex}(z, W) = s_{ex} \quad (2)$$

The output is scaled and is given by:

$$\mathbb{f}_X = \mathbb{f}_{scale}(s_{ex}, m_{se}) \quad (3)$$

To incorporate local descriptions of the visual semantics, local objects in the scene are detected using YOLO-V4 [46]. DropBlock regularization in YOLO-V4 supports better detection speed, accuracy, and mean average precision. The detected object features are flattened and passed through 3-fully connected layers resulting in a feature vector of dimension $[1 \times 256]$. Now, both the global and local visual encodings are concatenated as $\mathrm{F}_{vis} = \{\mathbb{f}_1, \mathbb{f}_2, \ldots . \mathbb{f}_N\}, \mathbb{f}_i \in \mathcal{R}^{\mathcal{D}}, i \in (1, N)$ and are further merged with fine-tuned GloVe [47] embedding-based text encodings given by $\mathcal{T}_{text}$. Unlike Word2Vec [48], GloVe embeddings highlight word co-occurrences to obtain the word embeddings. Now, the text and visual encodings are given as $\mathcal{X} = \{\mathrm{F}_{vis}, \mathcal{T}_{text}\}$, are combined and passed to the RFCG decoder.

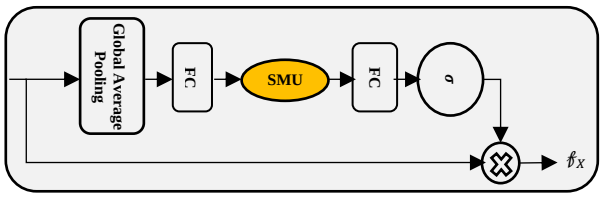


Fig. 4: SMU-activated SENet

(ii) RFCG-Decoder

This sub-section describes briefly the language generation process involved in the proposed RFCG. The language decoder is based on the Bi-LSTM network. The Bi-LSTM network solves well the problem of fixed sequence-to-sequence prediction. The combined text and visual embeddings $\mathcal{X} = \{\mathrm{F}_{vis}, \mathcal{T}_{text}\}$ are fed to the Bi-LSTM layer. The output generated from the Bi-LSTM network is activated with sigmoid activation. Also, the sigmoid-activated (σ) output is attended to using a soft attention mechanism. The soft attended features obtained are learned with the help of F_{vis}. By multiplying the related segmentation map with low weight, soft attention discredits unimportant parts. Therefore, high attention zones retain their original worth, whereas low attention areas approach zero. The process can be described mathematically as:

$$\mathbb{S}_i^t = \mathcal{W}_{\mathbb{S}} \tanh(\mathcal{W}_{\mathcal{X}} \mathcal{X}_i + \mathcal{W}_h h_{t-1} + \mathscr{b}_s) + \mathcal{W}_f \mathbb{f}_i \quad (4)$$

Where, are the $\mathcal{W}_{\mathbb{S}}$, $\mathcal{W}_{\mathcal{X}}$, $\mathcal{W}_{\mathbb{f}}$ $\mathcal{W}_h$ are the weights and $\mathscr{b}_s$ is the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 10:10:18 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3405573

bias factor. Also, $\mathbb{S}_i^t$ is the importance score and

$$\mathbb{S}_i^t = (\mathbb{S}_1^t, \mathbb{S}_2^t \ldots . \mathbb{S}_n^t)^T \quad (5)$$

Furthermore, the attention at time t is given by:

$$a_i^t = \frac{\exp(\mathbb{S}_i^t)}{\sum_{i=1}^{n} \exp(\mathbb{S}_i^t)} \quad (6)$$

For improved feature learning, the soft-attended and sigmoid-activated features are learned using LSTM layers. Further, the resulting output generates a word on the next time node using the MLP function that incorporates backpropagation through a time algorithm to update the parameters of the LSTM network. The objective function defined for the optimization of the proposed RFCG is given by:

$$\vartheta = arg \max_{\vartheta} \sum_{m,y} \sum_{t=0}^{N} \log p(y_t | m, \vartheta, y_1, y_2 \ldots y_{t-1}) \quad (7)$$

ϑ is the learnable parameter with m feature maps whose weight is $m \in \{m_1, m_2 \ldots m_t\}$. y is the sentence that describes the image well. Further, the cross-entropy loss is incorporated to minimize the loss function to maximize the probability of each correct word appearing. The cross entropy-loss function is given by:

$$\mathcal{L}(\vartheta) = -\sum_{t=1}^{N} \log\left(p_\theta(w_t^* | \mathbb{G}_{1:t-1}^*)\right) \quad (8)$$

where N is the number of words in a sentence; ϑ denotes all the parameters in the model; $\mathbb{G}_{1:t-1}^*$ is the Ground Truth (GT).

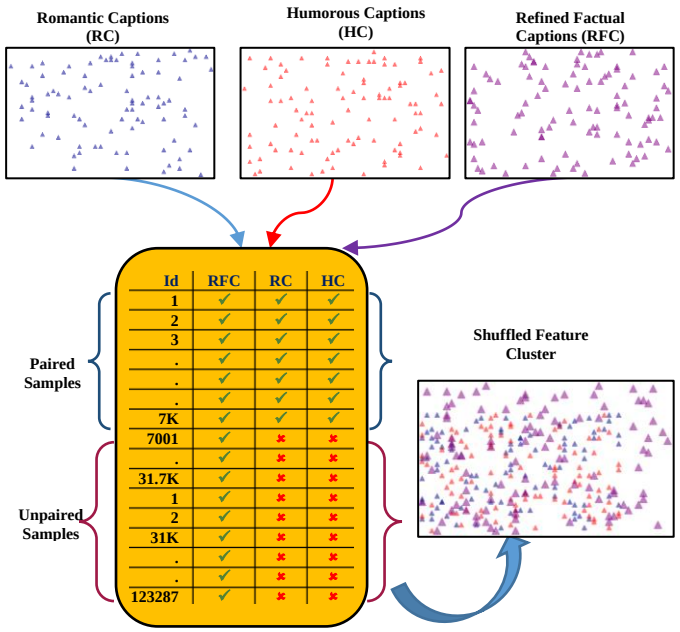


Fig. 5: Structure of Bag-of-Captions (BoC)

(iii) Bag-of-Captions (BoC)

The Refined Factual Captions (RFC) generated from the factual image captioning model, given as $RFC = \{RFC_1, RFC_2, RFC_3 \ldots RFC_n\}, n \in no.of\ sample\ images$ and the stylized captions- Romantic captions (RC) and Humorous captions (HC) combinedly define BoC, as shown in Fig. 5. It consists of 7K paired sample set as $\{RFC, RC, HC\}$ while the remaining samples include only RFC obtained from Flickr30K and MSCOCO datasets. Therefore, the BoC contains a set of paired captions and unpaired captions. All the samples are combined and shuffled and the shuffled feature cluster of captions is further split into training, validation, and test sets. Table I presents the details of training, testing, and validation splits. The contents of BoC are fed to SE-VAE-CSCG which are utilized for the generation of diverse and stylized descriptions of an image without directly depending on image features.

TABLE I
TEST, TRAIN & VALIDATION SPLITS FOR BOC

Datasets	Captions in BoC	Training Samples	Testing Samples	Validation Samples
Flickr30K FlickrStyle10K	RFC- 31,783 RC-7000 HC- 7000	32,048	6868	6867
MSCOCO FlickrStyle10K	RFC-1,23,287	1,15,287	11,000	11,000

B. SE-VAE-CSCG Model

It presents the Phase 2 of the proposed framework which is defined in two parts: (i) SE-VAE, and (ii) CSCG Module. The proposed modules generate stylized captions of an image using stylized embeddings based on variational autoencoder (SE-VAE) and controlled stylized caption generation. The architecture of the proposed method is depicted in Fig. 3.

(i) SE-VAE Module

The recent controlled text generation-based works [49] [50] [51] have explored the controlled text learning ability of Variational Auto-Encoder (VAE). In view of these works, the proposed work, attempts to introduce style-based caption representation for controlled text generations by defining modified Style-Embedding based Variational Autoencoder (SE-VAE), as depicted in Fig. 3. It receives paired and unpaired sample sets from BoC in the form of contextualized vector representation, derived from RoBERTa [52] by inferencing from both sentence and word embeddings. The VAE encoder constructs a latent distribution $\mathbb{N}(\mu, \sigma)$ using a mean and a variance vector. The latent distribution z is assumed to be a normal distribution whose loss function is defined as:

$$\mathcal{L}_{vae} = -\mathbb{E}_{q(z|x)}[\log p(x|z)] + \varphi.\mathbb{KL}(q(z|x)||p(z)) \quad (9)$$

Where the first term in the above equation represents the likelihood of the reconstruction of the original text x, while the second term is the KL-divergence between the latent distribution and standard normal distribution. Further, $p(z)$ is the prior for standard normal distribution, and $q(z|x)$ is the posterior distribution. Also, φ is the balancing parameter that learns the capacity between self-reconstruction and style features.

Previous works [41] [42] incorporated disentangling style attributes whereas the proposed method utilizes the style learnt from structured samples to generate target style representations for both unstructured and structured representations. It adjusts the style strength by simply adjusting the style weights. The embeddings are differentiated for different styles (romantic, humorous) using one hot vector encodings as Style-R and Style-H. It enhances the latent feature representations for both structures and unstructured latent vectors. Further, the style embeddings are represented by: $\mathbb{S} = \{s_1, s_2, \ldots, s_k\}, s_i \in \mathbb{R}^{k \times d}$, where $k = 2$ is the number of styles (Style-R and Style-H). These embeddings are randomly initialized and updated by minimizing the similarity between the style embeddings and latent fusion. The structure representation is depicted in Fig. 6, where the similarity is minimized using the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 10:10:18 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3405573

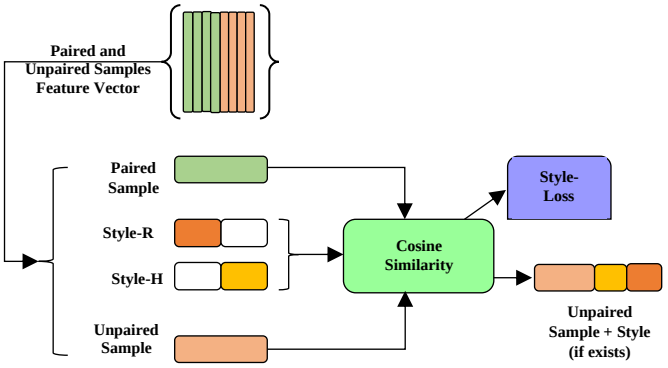


Fig. 6: Style-loss Calculation and Representation

cosine similarity with the assumption that style embedding is highly related to the latent features. The style loss for the same is defined as:

$$\mathcal{L}_{style} = -\sum_{i=1}^{k} \mathscr{b}_i \log\left(sigmoid\left(\cos\left(s_i, \mathcal{d}(z)\right)\right)\right) \quad (10)$$

In equation 7, the sigmoid function ensures the range of cosine similarity. $\mathscr{b}_i$=1 if the style given represents the style of the input sentence, otherwise $\mathscr{b}_i$=0. $\mathcal{d}(z)$ represents the stop gradient which is used to compute the style loss. Therefore, equation (9) is modified as:

$$\mathcal{L}_{\grave{vae}} = -\mathbb{E}_{q_{(Z|X)}}[\log p(x|z) + \mathcal{d}(\mathcal{s}_x)\,] + \varphi.\,\mathbb{KL}(q(z|x)||p(z)) \quad (11)$$

For sentence x, $\mathcal{s}_x$ represents its style embedding which is used as a constant vector. Hence, the total loss function defined for the proposed SE-VAE is:

$$\mathcal{L}'_{SE-VAE} = \delta_{vae}\mathcal{L}_{\grave{vae}} + \delta_{style}\mathcal{L}_{style} \quad (12)$$

Where, δ_{vae} and δ_{style} are the hyperparameters that balance the weights between VAE-loss and style loss.

(ii) CSCG Module

Controlled Stylized Caption Generation (CSCG) module, as shown in Fig. 3, leverages a generator and a discriminator structure. The output generated from the SE-VAE is represented in the form of unstructured representation $\mathcal{u}$ and structured representation $\mathcal{r}$. It trains the generator $\mathcal{G}$ that reconstructs the captions for generating plausible text. Further, the discriminator $\mathcal{D}$ enforces the generator to produce coherent attributes. The generator and the discriminator form a pair of collaborative learners and provide feedback signals to each other. Also, the collaborative optimization represents the wake-sleep algorithm [53].

The generator $\mathcal{G}$ is a Deep LSTM-RNN architecture that generates the final caption sequence in the form of tokens $\hat{x} = \{\hat{x}_1, \ldots, \hat{x}_T\}$ conditioned on $(\mathcal{u}, \mathcal{r})$, mathematically represented as:

$$\hat{x} \sim \mathcal{G}(\mathcal{u}, \mathcal{r}) = p_{\mathcal{G}}(\hat{x}|\mathcal{u}, \mathcal{r}) = \prod_t p(\hat{x}|\hat{x}^{<t}, \mathcal{u}, \mathcal{r}) \quad (13)$$

Where, $\widehat{x^{<t}}$ are the tokens preceding $\hat{x}$. Also, $\hat{x}$ is parametrized using the softmax function for time step t.

$$\hat{x} \sim softmax(o_t/\mathcal{T}) \quad (14)$$

$\mathcal{T} > 0$ is the temperature normally set to 1 and o_t is the logit vector.

The unstructured part incorporates standard Gaussian prior modeling for continuous variables whereas, the structured representation contains both continuous and discrete variables that encode different styles efficiently. Further, the SE-VAE is modified by combining the parameters of the encoder and generator. This modified SE-VAE loss is minimized to optimize the reconstructed captions, which is represented mathematically as:

$$\mathcal{L}_{SE-VAE} = -\mathbb{E}_{q_E(\mathcal{u}|X)q_{\mathcal{D}}(\mathcal{r}|X)}\left[\log p_{\mathcal{G}}(x|\mathcal{u}, \mathcal{r}) + \mathcal{d}(\mathcal{s}_x)\,\right] + \varphi.\,\mathbb{KL}(q_E(\mathcal{u}|x)||p(\mathcal{u})) \quad (15)$$

The distribution defined over $\mathcal{u}$ and $\mathcal{r}$ for the encoder and discriminator is mathematically given as:

$$\mathcal{u} \sim E(x) = q_E(\mathcal{u}|x) \quad (16)$$

$$\mathcal{D}(x) = q_{\mathcal{D}}(\mathcal{r}|x) \quad (17)$$

Further, these distributions are utilized as the output at the current step and the input to the next step along the sequence of decision-making. The resulting caption denoted by $\widetilde{\mathcal{G}_{\mathcal{T}}}(\mathcal{u}, \mathcal{r})$, is given as input to the discriminator that measures the fitness to the target attributes for structured and unstructured representation. The loss for the same is calculated as:

$$\mathcal{L}_{attr,\mathcal{r}} = \mathbb{E}_{p(\mathcal{u})p(\mathcal{r})}\left[\log q_{\mathcal{D}}(\mathcal{r}|\,\widetilde{\mathcal{G}_{\mathcal{T}}}(\mathcal{u}, \mathcal{r}))\right] \quad (18)$$

$$\mathcal{L}_{attr,\mathcal{u}} = \mathbb{E}_{p(\mathcal{u})p(\mathcal{r})}\left[\log q_E(\mathcal{u}|\,\widetilde{\mathcal{G}_{\mathcal{T}}}(\mathcal{u}, \mathcal{r}))\right] \quad (19)$$

Combining the equations, the final generator objective function is given as:

$$min(\mathcal{L}_{\mathcal{G}}) = \mathcal{L}_{SE-VAE} + \alpha\mathcal{L}_{attr,\mathcal{u}} + \beta\mathcal{L}_{attr,\mathcal{r}} \quad (20)$$

Where α and β are balancing parameters.

The discriminator is also a Deep-LSTM-based architecture that learned differently as compared to the SE-VAE encoder. The unstructured representation $\mathcal{u}$ is learned in an unsupervised manner while the structured representation $\mathcal{r}$ used labeled examples to entail a designated style (romantic or humorous). Therefore, efficient semi-supervised learning is defined for the discriminator. To learn the specified semantic meaning and style, a set of labeled examples $(\mathcal{X}_L = \{(x_L, \mathcal{r}_L)\})$ is used represented by:

$$\mathcal{L}_{disc} = \mathbb{E}_{\mathcal{X}_L}[\acute{\log}\, q_D(\mathcal{r}_L|x_L)] \quad (21)$$

Besides the conditional generator, $\mathcal{G}$ synthesizes noisy style-attribute pairs which used semi-supervised learning. To alleviate this issue minimum entropy regularization [54] is incorporated to provide robust model optimization. The resulting objective is given by:

$$\mathcal{L}_{\mathcal{D}-\mathcal{O}} = \mathbb{E}_{p_{\mathcal{G}}(\hat{x}|\mathcal{u}, \mathcal{r})p(\mathcal{u})p(\mathcal{r})}[log q_{\mathcal{D}}(\mathcal{r}|\hat{x}) + \rho\mathcal{H}\left(q_{\mathcal{D}}(\mathcal{r}'|\hat{x})\right)] \quad (22)$$

Where, $\mathcal{H}\left(q_{\mathcal{D}}(\mathcal{r}'|\hat{x})\right)$ is the empirical Shannon entropy of the distribution $q_{\mathcal{D}}$ and ρ is the balancing parameter. The final training discriminator objective is given by the following equation:

$$\min\left(\mathcal{L}_{\mathcal{D}}\right) = \mathcal{L}_{disc} + \theta\mathcal{L}_{\mathcal{D}-\mathcal{O}} \quad (23)$$

IV. RESULTS AND DISCUSSIONS

To evaluate the performance of the proposed framework MSCOCO, Flickr30K, and FlickrStyle10K datasets are used in the experimentations. The Flickr30K and MSOCOCO datasets are used for refined Factual Caption generation. To generate text-controlled stylized captions for any given image, style annotations for two different styles namely romantic and humorous styles are utilized from FlickrStyle10K [7]. For this dataset, only 7K style annotations are publicly available. This results in the collection of 7K paired samples and 24,783 from Flickr30K and 1,23,287 MSCOCO which are used for Phase II experiments.

A. Implementation Details

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 10:10:18 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3405573

In phase-1, the Refined Factual Caption Generation (RFCG) is trained with Adam [55] optimizer with a learning rate of $1e^{-5}$ for the encoder module whereas for the language decoder, the learning rate is set as $4e^{-4}$. Further, the batch size is set as 64 and RFCG is trained for 60 epochs. Also, to extract the text features, fine-tuned GloVe embeddings are utilized with dimensions for embeddings as 300. To evaluate the performance of the proposed RFCG, BLEU@N [56] and METEOR (M) [57] scores are used.

For the generation of stylized captions, in Phase-2, using the proposed SE-VAE-based controlled text generation, the Adam optimizer is utilized with a learning rate of 0.0005. The input to the SE-VAE is the embeddings extracted from RoBERTa with a batch size of 512 tokens. The dimensions for the latent features and style embeddings are $[1 \times 768]$. The model can generate sentences with a sentence length limit of ≤ 25. To evaluate the performance of generated stylized captions, Style Transfer Accuracy (cls) and Perplexity (ppl) of the proposed model is evaluated. Also, the relevancy of the captions generated is evaluated in terms of BLEU@N (B-1, B-3) [56] and METEOR (M) [57].

B. Experimental Results

This section discusses the results of the experiments conducted for two phases: (i) RFCG and, (ii) CSCG

(i) Refined Factual Captioning Generation (RFCG)

TABLE II
COMPARISON RESULTS ON FLICKR30K DATASET

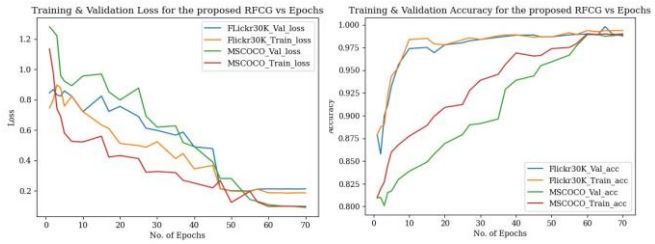
Model	B-1	B-2	B-3	B-4	M
LSTM [17]	66.3	42.3	27.7	18.3	-
G-LSTM [18]	64.6	44.6	30.5	20.6	17.9
Soft Attention [3]	66.7	43.4	28.8	19.1	18.5
Hard Attention [3]	66.9	43.9	29.6	19.9	18.5
Semantic Attention [2]	64.7	46.0	32.4	23.0	18.9
D-Ada [4]	66.7	48.6	32.1	22.4	21.4
Proposed RFCG	**69.6**	**49.4**	**32.1**	**23.3**	**21.6**

TABLE III
COMPARISON RESULTS ON MSCOCO DATASET

Model	B-1	B-2	B-3	B-4	M
LSTM-A [58]	75.4	-	-	35.2	26.9
VS-LSTM [59]	76.3	-	-	34.3	26.9
Up-Down [3]	77.2	-	-	36.2	27.0
RDN [19]	77.5	61.8	47.9	36.8	27.2
GCN-LSTM [60]	77.3	-	-	36.8	27.9
AoANet [20]	77.4	-	-	37.2	28.4
X-Transformer [25]	77.3	61.5	47.8	37.0	28.8
Fusion-LLM [26]	-	-	-	29.0	28.7
M2 [23]	**80.3**	-	-	39.1	29.2
Proposed RFCG	80.2	**62.6**	**48.4**	**39.4**	**29.4**

This section summarizes the performance of the proposed RFCG architecture on the Flickr30K and MSCOCO Karapathy split datasets. Table II and Table III presents the comparison of the proposed RFCG based generated refined captions with other state-of-the-art on Flickr30K and MSCOCO datasets. Compared with the other state-of-the-art in Table II, the proposed method provides a significant improvement in terms of B-1, B-2, B-3, B-4, and METEOR. The key reason for consistently improved performance of the proposed RFCG model is that it leverages the benefit of rich Yolo-V4 based region specific local features and Inception-V3 based global visual features while merging with GloVe embeddings-based text features in comparison to raw VGG 16 based encoders [4], and CNN based encoders [2] visual features. In addition to this the proposed RFCG module incorporates Bi-LSTM and rich visual features enabled soft attention-based language decoder which provides further enhancements in the generating refined captions when compared with hard and soft attention [3]. Also, From Table-III it is evident that the our proposed RFCG module provided significant results when compared with many recent state-of-the-art [20] [25] [58, 59, 60]. Also, [26] LLM-based fusion reports very low scores as fusion of captions collapse into a caption that does not offer more detail. To overcome this the proposed framework generates more detailed factual captions which are further fused with style-based captions for the generation of stylized image captions.

Further, when compared with [23] the model provides improvements in terms of B-2, B-3, B-4, and M only. Fig. 7(a) and Fig. 7(b) present the validation and test accuracy and loss curves for the proposed RFCG model. From this, it is evident that with the increase in the number of epochs, the accuracy of the proposed RFCG model increases while the losses decrease. It is evident from the figures that after 60 epochs the performance of the proposed RFCG module does not provide significant improvements in the results. Hence, early stopping criteria is used to stop the training after receiving consistent results for next 10 epochs. Fig. 8 presents the qualitative results of the proposed RFCG. It is observed that the proposed RFCG expresses the relationships between objects, attributes, and scenes with more refined syntactic and semantic descriptions, over base LSTM and other models [3, 4, 25].



(a) **(b)**

Fig. 7: Training & Validation (a) Loss, (b) Accuracy for the proposed RFCG

TABLE IV
COMPARISON RESULTS ON FLICKRSTYLE10K DATASET
TEST CASE 1: 24,783 UNPAIRED SAMPLES FROM FLICKR 30K; **TEST CASE 2:** 1,23,287 UNPAIRED SAMPLES FROM MSCOCO

Style	Model	B-1	B-3	M	cls	ppl
Romantic	StyleNet [7]	13.3	1.5	4.5	57.1	6.9
	MSCap [28]	17.0	2.0	5.4	91.3	-
	SF-LSTM [6]	27.8	8.2	11.2	-	-
	Detach and attach [30]	24.3	-	-	82.4	-
	Wu et al. [31]	25.4	5.7	9.2	-	-
	SAN [11]	28.3	8.7	11.5	90.9	9.1
	SE-VAE-CSCG (Test Case-1)	**30.2**	**9.4**	**12.5**	**91.6**	**9.7**
	SE-VAE-CSCG (Test Case-2)	**32.8**	**12.2**	**13.1**	**94.1**	**12.1**
Humorous	StyleNet [7]	13.4	0.9	4.3	42.5	7.3
	MSCap [28]	16.3	1.9	5.3	88.7	-
	SF-LSTM [6]	27.4	8.5	11.0	-	-
	Detach and attach [30]	23.0	-	-	89.2	-
	Wu et al. [31]	27.2	5.9	9.0	-	-
	SAN [11]	27.6	8.1	11.2	87.8	8.4
	SE-VAE-CSCG (Test Case-1)	**29.7**	**9.0**	**11.6**	**89.7**	**9.3**
	SE-VAE-CSCG (Test Case-2)	**31.2**	**10.6**	**12.0**	**92.4**	**11.6**

(ii) SE-VAE-CSCG

For this phase, we define two test cases for our experiment: i) Test Case 1 and ii) Test Case 2. Test Case 1 utilizes 24,783

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 10:10:18 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3405573

Fig. 8: Refined Captions Generated by Proposed RFCG Model on Flickr30K and MSCOCO datasets.

Fig. 9: Stylized Romantic and Humorous Captions Generated Using Controlled Stylized Caption Generation (CSCG) (a) With Flicr30K unpaired samples (24,783) and (b) With MSCOCO unpaired samples (1,23,287 samples)

unpaired samples from Flickr 30K whereas Test Case 2 utilizes 1,23,287 unpaired samples from MSCOCO. Also, both the test cases contain 7K paired samples. Table IV presents the comparison of the proposed SE-VAE CSCG model for both the test cases with other state-of-the-art in terms of BLEU-1, BLEU-3, and METEOR (M) scores with style accuracy (cls) and perplexity (ppl). StyleNet [6] and SF-LSTM [7] incorporate factored LSTM and style-factual LSTM decoders for end-to-end learning of stylized romantic and humorous captions. The works [11] [28] [30] [31] attempted multi-style-based caption learning using unpaired data but the proposed SE-VAE-CSCG captures the unstructured and stylized disentangled representations better than [11] [28], using the concept of controlled styled SE-VAE, with the significant rise of B-1 score value by an amount of 3.9, 15.8, 8.5 and 7.4 respectively. From Table IV we can infer that Test Case 2 provides significant results for all the evaluation metrics. Hence, increasing the number of unpaired samples confirms efficient learning of linguistic styles in the final encodings Furthermore, it can be observed from the qualitative results of the proposed SE-VAE-CSCG, as reported in Fig. 9, that the image content is described with the usage of the correct style i.e., romantic and humorous with more realism, fluency, grammatical correctness and diversity.

A. Ablation Studies

An ablation study is carried out to exhibit the influence of (i) RFCG module and (ii) SE-VAE-CSCG module (iii) number of unpaired captions fed to the BoC from MSCOCO dataset on the performance of the proposed framework.

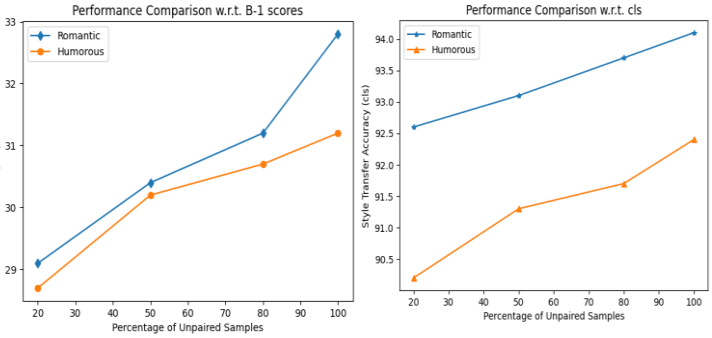


Fig. 10: Ablation study comparison results for percentage of number of samples utilized for generation of romantic and humorous captions

TABLE V
Ablation Study Results for Proposed RFCG Module for Flickr30K Dataset

Encoder Feature Representation	B-1	B-2	B-3	B-4	M
VGG-16	62.0	43.4	26.1	19.2	17.8
Inception-V3	64.7	45.2	26.2	19.5	18.9
R-CNN	65.8	45.7	27.3	20.1	19.2
Yolo-V4	65.9	45.7	28.1	20.7	19.3
Inception-V3 + Yolo-V4	**69.6**	**49.4**	**32.1**	**23.3**	**21.6**

i) In RFCG module the encoder utilizes efficient Inception-V3 + Yolo-V4 visual features. The key reason to use Inception-V3 + Yolo-V4 based features, is highlighted in the ablation study reported in Table V, for Flickr30K dataset. It confirms that RFCG module performs superior for Inception-V3 + Yolo-V4 based visual features over VGG-16, Inception V3, R-CNN, Yolo-V4, with a prominent rise in B-1, B-2, B-3, B-4 and M scores.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 10:10:18 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3405573

TABLE VI
Ablation Study Results for Proposed SE-VAE-CSCG Module

Style	Case	Basic-VAE	VAE	SE	CSCG	B-1	B-3	M	cls	ppl
Romantic	Test Case 1	✓	×	×	✓	18.7	5.4	6.2	66.7	6.5
		×	✓	×	✓	27.2	8.2	10.5	86.4	7.9
		×	✓	✓	✓	**30.2**	**9.4**	**12.5**	**91.6**	**9.7**
	Test Case 2	✓	×	×	✓	18.9	6.1	6.4	68.3	6.8
		×	✓	×	✓	28.1	8.6	10.8	87.9	7.9
		×	✓	✓	✓	**31.7**	**11.1**	**14.2**	**93.7**	**11.3**
Humorous	Test Case 1	✓	×	×	✓	18.2	5.1	6.1	62.1	6.2
		×	✓	×	✓	26.9	8.1	10.2	85.7	7.2
		×	✓	✓	✓	**29.7**	**9.0**	**11.6**	**89.7**	**9.3**
	Test Case 2	✓	×	×	✓	18.8	5.4	6.3	67.1	7.1
		×	✓	×	✓	27.4	8.7	10.9	87.0	7.5
		×	✓	✓	✓	**30.3**	**10.1**	**11.9**	**91.4**	**10.6**

ii) The ablation study results for proposed SE-VAE-CSCG module are shown in Table VI. Where it can be observed that the style-based captions are generated using three different models Basic-VAE + CSCG, VAE + CSCG, and SE-VAE + CSCG respectively. It is observed that there is a significant increase in the B-1, B-2, M, cls, and ppl scores for the proposed SE-VAE-CSCG module. The scores obtained for both Test Case 1 and 2 makes it evident that the proposed SE-VAE-CSCG module generates stylized descriptions with favorable accuracy, and perplexity trade-offs by efficiently lifting the word-level knowledge to sentence-level knowledge and learning disentangled representations.

iii) Further, ablation is also carried out to study the influence of number of unpaired captions fed to the BoC from MSCOCO dataset to exhibit the generalizability and independence of the proposed framework over limited no of structured captions. Initially, for MSCOCO and FlickrStyle10K dataset we utilized 7K paired samples and 20% unpaired samples from MSCOCO. With increase in the number of samples from 20% to 50%, 80%, and 100% samples a prominent increase in values of B-1 and cls is observed in Fig. 10.

V. Conclusion

This paper presents a novel style-based caption generation framework that generates style-controlled descriptions of images in two phases: (i) Refined Factual Caption Generation (RFCG), and (ii) Controlled Style Caption Generation, SE-VAE-CSCG. The proposed framework generates realistic and stylized descriptions of images with controlled text generation. The proposed RFCG generates meaningful human-independent (unbiased) refined captions by merging both visual features (global features and object-level local features) and textual features. The presented SE-VAE-CSCG module exhibits independence on the number of structured samples, as it delivers refined stylized captions even for majority of unstructured samples projected in terms of improved style accuracy.

We also present a novel Bag of Captions (BoC) concept which is a collection of both paired and unpaired samples of captions. This helps to learn disentangled representations by lifting the word-level knowledge to sentence-level knowledge. Further, an ablation study is also conducted to support the experiments. In future work, the focus can be extended to large-scale stylized datasets incorporating different styles.

References

[1] D. Sharma, C. Dhiman and D. Kumar, "Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey," *Expert Systems with Applications*, vol. 221, p. 119773, 2023.

[2] Q. You, H. Jin, Z. Wang, C. Fang and J. Luo, "Image captioning with semantic attention," in *IEEE Conference on Computer Vision and Pattern Recognition*, Caesars Palace, 2016.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 2018.

[4] Z. Deng, Z. Jiang, R. Lan, W. Huang and X. Luo, "Image captioning using DenseNet network and Adaptive Attention," *Signal Processing: Image Communication*, vol. 85, p. 115836, 2020.

[5] J. Yu , J. Li, . Z. Yu and Q. Huang, "Multimodal Transformer With Multi-View Visual Representation for Image Captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4467-4480, 2020.

[6] T. Chen, Z. Zhang, Q. You, C. Fang, Z. Wang, H. Jin and J. Luo, ""Factual" or "Emotional": Stylized Image Captioning with Adaptive Learning and Attention," in *European Conference on Computer Vision*, Munich, Germany, 2018.

[7] C. Gan, Z. Gan, X. He and J. Gao, "Stylenet: Generating attractive visual captions with styles," in *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, 2017.

[8] T. Li, Y. Hu and X. Wu, "Image Captioning with Inherent Sentiment," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, Shenzhen, China, 2021.

[9] A. Mathews, L. Xie and X. He, "SentiCap: generating image descriptions with sentiments," in *AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[10] K. Shuster, S. Humeau, H. Hu, A. Bordes and J. Weston, "Engaging Image Captioning via Personality," in *arXiv:1810.10665*, 2018.

[11] G. Li, Y. Zhai, Z. Lin and Y. Zhang, "Similar Scenes arouse Similar Emotions: Parallel Data Augmentation for Stylized Image Captioning," in *ACM Multimedia Conference*, Virtual Event, China, 2021.

[12] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar and C. Lawrence, "Microsoft coco: Common objects in context," in *European Confrence on Computer Vision*, Zurich, Switzerland, 2014.

[13] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67-78, 2014.

[14] Y. Yang, C. L. Teo, H. Daume and Y. Aloimono, "Corpus-guided sentence generation of natural images," in *Empirical Methods in Natural Language Processing*, Edinburgh United Kingdom, 2011.

[15] M. Hodosh, P. Young and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics,"

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 10:10:18 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3405573

Journal of Artificial Intelligence Research, vol. 47, pp. 853-899, August 2013.

[16] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," in *arXiv:1412.6632*, 2015.

[17] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , Boston, USA, 2015.

[18] X. Jia, E. Gavves, B. Fernando and T. Tuytelaars, "Guiding the long-short term memory model for image caption Generation," in *arXiv:1509.04942*, 2015.

[19] L. Ke, W. Pei, R. Li, X. Shen and Y.-W. Tai, Reflective Decoding Network for Image Captioning, arXiv:1908.11824, 2019.

[20] L. Huang, . W. Wang, J. Chen and X.-Y. Wei, "Attention on Attention for Image Captioning," in *arXiv:1908.06954*, 2019.

[21] J. Lu, C. Xiong, D. Parikh and R. Socher, "Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning," in *arXiv:1612.01887v2*, 2017.

[22] V.-K. Vo-Ho, Q.-A. Luong, D.-T. Nguyen, M.-K. Tran and M.-T. Tran, "A Smart System for Text-Lifelog Generation from Wearable Cameras in Smart Environment Using Concept-Augmented Image Captioning with Modified Beam Search Strategy," *Applied Sciences*, vol. 9, no. 9, p. 1886, 2019.

[23] M. Cornia, M. Stefanini, L. Baraldi and R. Cucchiara, Meshed-Memory Transformer for Image Captioning, arXiv:1912.08226, 2020.

[24] G. Li, L. Zhu, P. Liu and Y. Yang, "Entangled transformer for image captioning," in *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, 2019.

[25] Y. Pan, T. Yao, Y. Li and T. Mei, "X-Linear Attention Networks for Image Captioning," in *arXiv:2003.14080v1*, 2020.

[26] S. Bianco, L. Celona, M. Donzella and P. Napoletano, "Improving Image Captioning Descriptiveness by Ranking and LLM-based Fusion," in *arXiv:2306.11593*, 2023.

[27] N. Rotstein, D. Bensaïd, S. Brody, R. Ganz and R. Kimmel, "FuseCap: Leveraging Large Language Models for Enriched Fused Image Captions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024.

[28] L. Guo, J. Liu, P. Yao, J. Li and H. Lu, "MSap: Multi-Style Image Captioning With Unpaired Stylized Text," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] A. P. Mathews, L. Xie and X. He, "SentiCap: Generating Image Descriptions with Sentiments," in *AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, 2016.

[30] Y. Tan, Z. Lin, P. Fu, L. Wang, Y. Cao and W. Wang, "Detach and Attach: Stylized Image Captioning without Paired Stylized Dataset," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.

[31] X. Wu, W. Zhao and J. Luo, "Learning Cooperative Neural Modules for Stylized Image Captioning," *International Journal of Computer Vision*, vol. 130, p. 2305–2320, 2022.

[32] D. Yang, H. Chen, X. Hou, T. Ge, Y. Jiang and Q. Jin, "Visual Captioning at Will: Describing Images and Videos Guided by a Few Stylized Sentences," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

[33] S. Gu, C. Clark and A. Kembhavi, "I can't believe there's no images! Learning Visual Tasks Using Only Language Supervision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[34] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov and E. P. Xing, "On unifying deep generative models," in *arXiv:1706.00550*, 2017.

[35] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2014.

[36] A. B. L. Larsen, S. K. Sønderby, H. Larochelle and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International Conference on Machine Learning (ICML)*, New York NY USA, 2016.

[37] Z. Yang, Z. Hu, R. Salakhutdinov and T. Berg-Kirkpatrick, "Improved variational autoencoders for text modeling using dilated convolutions," in *International Conference on Machine Learning*, Sydney, Australia, 2017.

[38] L. Yu, W. Zhang, J. Wang and Y. Yu, "SeqGAN: sequence generative adversarial nets with policy gradient," in *AAAI Conference on Artificial Intelligence*, San Francisco California USA, 2017.

[39] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever and P. Abbeel, "InfoGAN: interpretable representation learning by information maximizing generative adversarial nets," in *International Conference on Neural Information Processing Systems*, Barcelona Spain, 2016.

[40] N. Siddharth, B. Paige, J.-W. v. d. Meent, A. Desmaison, N. D. Goodman, P. Kohli, F. Wood and P. H. Torr, "Learning Disentangled Representations with Semi-Supervised Deep Generative Models," in *31st Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017.

[41] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov and E. P. Xing, "Toward Controlled Generation of Text," in *International Conference on Machine Learning (PMLR)*, Sydney, Australia, 2017.

[42] F. Betti, G. Ramponi and M. Piccardi, "Controlled Text Generation with Adversarial Learning," in *13th International Conference on Natural Language Generation*, Dublin, Ireland, 2020.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[44] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," in *arXiv:1709.01507*, 2019.

[45] K. Biswas, S. Kumar, S. Banerjee and A. K. Pandey, "SMU: SMOOTH ACTIVATION FUNCTION FOR DEEP NETWORKS USING SMOOTHING MAXIMUM TECHNIQUE," in *arXiv:2111.04682v2*, 2022.

[46] A. Bochkovskiy, C.-Y. Wang and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," in *arXiv:2004.10934v1*, 2020.

[47] J. Pennington, R. Socher and C. Manning, "GloVe: Global Vectors for Word Representation," in *Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014.

[48] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *arXiv:1301.3781*, 2013.

[49] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz and S. Bengio, "Generating Sentences from a Continuous Space," in *SIGNLL Conference on Computational Natural Language Learning*, Berlin, Germany, 2016.

[50] V. John, L. Mou, H. Bahuleyan and O. Vechtomova, "Disentangled Representation Learning for Non-Parallel Text Style Transfer," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019.

[51] H. Xu, S. Lu, Z. Sun, C. Ma and C. Guo, "VAE based Text Style Transfer with Pivot Words Enhancement Learning," in *arXiv:2112.03154v1*, 2021.

[52] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," in *arXiv:1907.11692*, 2019.

[53] G. E. Hinton, P. Dayan, B. J. Frey and R. M. Neal, "The "Wake-Sleep" Algorithm for Unsupervised Neural Networks," 1995.

[54] R. Scott, H. Lee, D. Anguelov, C. Szegedy, D. Erhan and A. Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," in *arXiv:1412.6596*, 2014.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic," in *arXiv:1412.6980*, 2014.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 10:10:18 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Cognitive and Developmental Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2024.3405573

[56] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, ""Bleu: a method for automatic evaluation of machine," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002.

[57] S. Banerjee and A. Lavie, "Meteor: An automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[58] T. Yao, Y. Pan, Y. Li, Z. Qiu and T. Mei, "Boosting Image Captioning with Attributes," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[59] N. Li and Z. Chen, "Image cationing with visual-semantic LSTM," in *Proceedings of the 27th International Joint Conference on Artificial Intellegence*, 2018.

[60] T. Yao, Y. Pan, Y. Li and T. Mei, "Exploring visual relationship for image captioning," in *ECCV*, 2018.

Dhruv Sharma, received his B.Tech degree from GGSIPU University in 2015 and M.Tech. degree from Ambedkar Institute of Advanced Communication Technology & Research, New Delhi, India, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electronics and Communication Engineering from Delhi Technological University, New Delhi, India. His current research interest includes machine learning, deep learning, computer vision, natural language processing and image captioning.

Chhavi Dhiman (M'16) received her B.Tech. degree from Indira Gandhi Delhi Technical University for Women (IGDTUW), Delhi, India, in 2011, M.Tech. and Ph.D. degree from Delhi Technological University (DTU), Delhi, India, in 2014 and 2019 respectively. She is currently working as Assistant Professor, in the Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India. Her research interest includes Machine Learning, Deep Learning, Pattern Recognition, Computer Vision, Human Action Identification and Classification. She is a reviewer of various Journals/Transactions of IEEE, Elsevier, and Springer. She has been awarded "Premium Research Award" by Delhi Technological University, Delhi, India in 2021.

Dinesh Kumar received his B.Tech. (ECE) and M.Tech.(ECE) (Gold Medalist) form National Institute of Technology (NIT) Kurukshetra (erstwhile REC Kurukshetra) and received Ph.D. degree in the field of Biometrics (Face Recognition) from GGS Indraprastha University, Delhi. He has more than 30 years of teaching and research experience and currently working as Professor in the Electronics and Communication Engineering Department of Delhi Technological University, Delhi. He has research publications in National/International Journals and Conferences including IEEE, Elsevier, Springer, Wiley. His current research interests include image processing, biometrics, soft computing, clustering, biomedical imaging, machine learning, image captioning. He is on the panel of reviewers for IEEE, Elsevier and Springer journals/Transactions.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 10:10:18 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

I. J. Computer Network and Information Security, **2024, 3, 26-36**
Published Online on June 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijcnis.2024.03.03

Cooperative MAC Protocol based on Best Data Rate (CMAC-DR)

Nagesh R.*
VTU, Belgavi, India
E-mail: Nagesh.r@gmail.com
ORCID iD: https://orcid.org/0000-0002-7867-4895
*Corresponding author

M. N. Sree Rangaraju
E&C, BIT, Bengaluru, India
E-mail: mnsrr@rediffmail.com

N. R. Kushar Gowda
E&C, GSKSJTI, Bengaluru, India
E-mail: kushargowda17@gmail.com
ORCID iD: https://orcid.org/0009-0007-9333-1034

Vinayak Shekharappa Antin
FEV Creltech GmbH, Lindau, Germany
E-mail: antinvinayak@gmail.com
ORCID iD: https://orcid.org/0009-0009-1697-2000

Received: 17 March 2021; Revised: 06 August 2021; Accepted: 23 April 2024; Published: 08 June 2024

Abstract: As wireless signals are broadcast in nature, which implies that, a broadcast communication purposive to a predetermined destination may be received by a non-intended intermediate station. Cooperative transference, which employ interposed stations to pass on the eavesdropped data to attain the contrast gains, has a substantial capability to revamp the channeling effectiveness in wireless systems. In this it is evident that having cooperation amid stations in a wireless system can accomplish higher throughput with enhanced network lifetime. Proffered work bestows a model for medium access layer called Cooperative MAC protocol based on optimal Data Rate (CMAC-DR). In the proffered work, stations with more data rate aid stations having lesser data rate in their communication by redirecting their congestion. In CMAC-DR model, utilizing the conveyance of eavesdropped information, potential helper stations with more data rate Send out Helper Ready to Send (HRTS), the stations with less data rate maintains a table, called Co-op table of potential helper stations, that can aid in its transmissions. During communication, the source station with low data rate chooses either transmitting by the way of a helper station, so that it lowers the end to end transference delay and increases the throughput or opt only direct transmission, if no potential helper is found or if CMAC-DR becomes an overhead. By analyzing varied simulated scenarios, CMAC-DR evaluates the elevation in the overall network lifetime, throughput and minimization of delay. The CMAC-DR protocol is unambiguous and in accordant with legacy 802.11 also when compared to this, we find improved performance in terms of delay throughput and network lifetime since data rate is considered as relay selection condition.

Index Terms: Wireless Adhoc Systems, Cooperative Transference, Medium Access Control, Data Rate, Throughput, CoopTable, Quality of Service(QoS).

1. Introduction

In contemporary time, wireless systems are advancing into a versatile devices beside immense data executions done by them [1]. Aforesaid accomplishments need fast connectedness and sturdy fault preventiveness [2,3]. Those necessities, besides enormous growth of wireless networks with constrained spectrum availability has evolved into momentous complications like extent crisis and more interventions. This circumstance necessitates an advance in the direction of origination of a modish wireless strategy that can facilitate a more methodical use of the accessible spectrum. While

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

unfolding strategies like "Multiple-Input-Multiple-Output" (MIMO) system, commences to advance the spectrum capability, but its utilization is constricted as its dimensions, price and the power for the compact wireless systems poses certain restrictions [4].

An auxiliary technique called cooperative transference guarantees to furnish some merits of the MIMO with certain restraints. The major objective was to introduce a best relay selection scheme for a multi hop communication between source and destination, there by improvising network performance. Cooperative transference cites in to the synergetic refining and relaying of an eavesdropped data by the stations which are encompassing source station. The perception of cooperative transference is to fully leverage the disseminating complexion of the wireless channel and to create a geographical diversification, specifically transfer variance, thereby attaining immense boost in system robustness, capability, delay, notable depletion in interventions, and extension of coverage scale [5]. The preliminaries of cooperative transference are found at the physical layer, although, the conception of cooperative transference is obtainable in varied dispositions at higher layer protocols. To be able to retrieve the physical layer data and to have swift flexibleness to sustained potency, it is legitimate to instigate the concept of cooperative transference in the layer which is immediate aloft the physical layer, viz. the medium access control (MAC) sublayer.

In cooperative transmission, numerous stations in wireless adhoc system function in adjoint to behave as an effective antenna assembly. Cooperative transference at the physical layer utilizes the broadcast character of the wireless system and eavesdropped information to enhance the effectiveness [6]. The stereotypical MAC models have long-drawn treated this attribute as a complication, preferably considering it as an alternative to overcome many bottlenecks. The methodology of cooperation employs this feature, and fabricate a state-of-the-art MAC model in wireless adhoc systems. The Cooperative MAC protocol based on optimal Data Rate (CMAC-DR) model for IEEE 802.11 wireless adhoc networks avail oneself of both the transfer characteristic of the wireless channel and cooperative distinctiveness. Contemporary 802.11 MAC models are primarily focused on throughput and capability under overloaded circumstances, particularly when every station contains a frame in standby condition, all the time [7]. The studies on "saturated environments" are required to obtain understanding on functioning of the protocol. Moreover, the overloaded consideration is inappropriate for present-day utilizations viz. audio and video streaming.

In CMAC-DR model, we proffer the implementation of cooperation at MAC sublayer by considering a wireless adhoc systems, in which there are potential helper/ Relay stations (RS) through which the data frame from the source station might also get forwarded to the destination station and create a cooperative scenario. We study the performance, by carrying out a broad lay of simulations rooted on the scheme features as these are enumerated below:

- Throughput performance under heavy load.
- Network lifetime.
- Energy consumption.
- Delay performance.

The balance of the paper is systemized as follow: In section-II, analogous studies is chronicled. Section-III deals with the working of the proffered CMAC-DR model. Section-IV deals with the simulation results. The analysis of the performance parameter of proffered model is established in section-V, Section VI provides the interpretation of the proffered model and the possible subsequent considerations.

2. Related Works

According to "IEEE-802.11-Distributed Coordination" Function (DCF)[1], in the wake of a source station recognizing a channel as inactive for a span of Distributed Inter-Frame Space(DIFS), source station will back down from sensing the channel for a duration chosen between 0 and it's "Contention Window"(CW) [8-9]. Subsequently as the timer finishes, the frame exchanging procedure is initiated. Any intermediate stations eavesdropping any of the control frames (RTS/CTS) obtains the duration info from the frame header, and places its Network allocation vector(NAV). Now this is referred as an exposed station problem in the IEEE 802.11 systems in which all the intermediate stations which are receiving the control frames which are not intending to them [10], will be silenced for the duration specified in the frames. Though certain stations communication of other stations will not be interfered with the current transference, it will not be allowed to do so. Thereby overall network efficiency, throughput, and other Quality of Service(QoS) parameters are degraded.

CoopMAC[2], Cooperative MAC protocol for adhoc wireless systems. This protocol is formed on an aim of involving an intermediate station in an ongoing transmission of the source and destination stations. The intermediate station retransmits the information eavesdropped from the source to the destination [11]. Consequently, lessening the transmission duration for the congestion being managed. In CoopMAC, the relay selection function is done by the source station, it chooses the best intermediate station which dispenses the least transference delay to destination. Substantial simulations depict that CoopMAC enhances the throughput and delay efficiently. For each simulation, stations were arbitrarily localized in an area of radius 350 m. Increasing the network size, coverage range, throughput, reducing source to destination delay and including mobility in accordance with real time scenario was considered as a subsequent study.

DELCMAC[12], The intent of the diversified energy location based cooperative MAC(DEL-CMAC) model is to

better the efficiency of mobile adhoc networks MANETs with respect to overall network lifespan and energy. The energy utilization model considers the energy utilization of transceiver circuits along with the transmit amplifier [13]. A diversified utilization-focused optimum relay selection strategy is used to select relay on the basis of locale data and energy remnant. Also, for the purpose of enhancing the structural reuse, a resourceful NAV setting is considered to deal with the varying transference power of the source and intermediate stations. In this paper, considering hidden terminal issue, increased network capacity with more mobility, minimizing the throughput and delay degradation is kept as a future work.

The rDCF model presented in [14] chooses an agile process by publicizing the capability of each intermediate station to help through Hello frames. Nonetheless in both CoopMAC and rDCF, the helper node forwards the frame to the destination, immediately after the reception from source station. But the relay station in [15] transmits the frame only if there is no reception of an ACK from destination indicating that it is unable to decipher the frame of the direct transference. Persistent RCSMA reported in [14] utilizes the function of a diversified and cooperation through automatic retransmission request (ARQ) model. In this model, if the destination station deciphers a data frame containing errors through direct transference from the source station, it requests a transference from any of the intermediate stations which has eavesdropped the original transference.

In Coop MACA protocol [15], it considers packet aggregation and distributed contention based relay selection algorithm to furnish a robust alterations instantaneously for any changes in the network architecture specially for a general mobile situations. The author of the EMR model proposes the consideration of a priority-mapping scheme of existing throughput and unify the information into the control frames in order to choose the optimum relay node.

3. Proffered CMAC-DR Model

For cooperative transference, we mainly focus on a wireless adhoc system, where the stations can enhance their effective quality of service(QoS) by cooperation [16]. In which, every wireless station is presumed to act as a source station as well as a cooperative agent for another station as shown in figure 1. In cooperative wireless system, source station initially transfers their frames to the intermediate helper station, then the helper station proceeds onto forwarding its received data information to the destination station.

The CMAC-DR model is built upon the IEEE-802.11-Distributed-Coordination Function(DCF [17]. The transference powers for all the stations are assumed to be unvaried. Request to send(RTS) and clear to send(CTS) frames can be eavesdropped by other intermediate stations besides the source and destination stations. If source node has a MAC protocol data unit (MPDU) to transmit, it has an option to transfer straight to the destination, or to employ an intermediate helper station for transference [4]. Therefore, in CMAC-DR model, initially as in legacy mode of 802.11, RTS frame will be broadcasted but in which the RTS frame is modified in such a way that it also contains data rate of source station as shown in figure.2. Further including destination even all other stations overhear the RTS frame and gets the data rate of source station. Subsequently modified CTS frame as shown in figure.3. will be sent by destination station by including its data rate.

The potential helper station which has eavesdropped both RTS and CTS frames compares the data rate of the source with its own. If it is of higher data rate than that of the source station, then it enables its cooperation mode and broadcasts HRTS(Helper ready to send) frame in which the data rate of potential helper station will be incorporated in the frame structure (Figure.4.) [18]. Here potential helper station refers to the station which is in the coverage region of both source and destination stations with more data rate than that of the source and also not involving itself with any of communications with any other stations and also not being allocated with NAV from any nearby transmissions [19]. Thereby source station receives the HRTS frames of all the probable helper stations and maintains a Coop table in its cache memory. Further it compares and chooses the best relay station which has the highest data rate to transmit the data frame. As a consequence, increase in the channel efficiency with higher throughput and less delay is achieved [7]. If the source station does not receive an HRTS within its timeout, it will then initiate a straight transference to the destination. An acknowledgement(ACK) is sent to for a correct acceptance, regardless of the event of frame forwarded by an intermediate helper station, or is directly transferred from the source.

It is important to take into account that each station gathers and revises the information about the potential helper station at one's disposal and their respective data rates only when it is required [3]. The CMAC-DR model transacts this issue through cooperation(Co-op) table. The Co-op table will be formed by source node (Table1), for keeping the entry of an information with regards to a potential helper station address viz., 48-bit MAC address and its data rate of direct transference between the potential helper station and destination.

By the virtue of broadcast property of a wireless channel, the destination station will obtain the data frame by both source station and relay station [10]. In the event of destination correlating these two copies to decipher the original information efficiently, then cooperative diversity is completely leveraged. Similar modification is done on the Data packet (Figure.5), so that it get forwarded through the potential relay as per the proffered CMAC-DR model. The Data frame is transmitted by source (SRC) station and from the other direction via a helper station by a process of address exchange mechanism at the helper station i.e., Address1 field in the data frame is replenished with the destination station address and broadcasted with its data rate [20]. The destination node or the station which has sent clear to send (CTS) frame, after successfully decoding the same data frames which have arrived in two subsequent time slots, replies with an

ACK frame so that it acknowledges the affluent acceptance of the data frame either by direct transmission from source side or through the CMAC-DR model. Figure.6. represents the complete action flow of the CMAC-DR model.

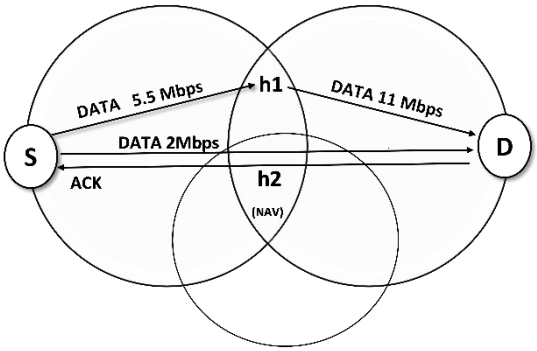


Fig.1. Cooperative scenario

Octets: 2	2	6	6	2	4
Frame control	Duration	Address1 (DA)	Address2 (SA)	Data Rate	FCS

Fig.2. Modified RTS control frame

Octets: 2	2	6	2	4
Frame Control	Duration	Address1 (DA)	Data Rate	FCS

Fig.3. Modified CTS control frame

Octets: 2	2	6	6	2	4
Frame control	Duration	Address1 (DA)	Address2 (SA)	Data Rate	FCS

Fig.4. Frame structure of HRTS control frame

Frame Control	Duration	Address 1 (Helper)	Address 2 (SRC)	Address 3(DST)	Sequence Control	Address 4	Frame body	FCS

Fig.5. Data frame format

Table 1. Fields of CMAC-DR model Co-op table

Helper MAC Address	Data Rate(Mbps)
00:00:00:00:00:02	5.5
00:00:00:00:00:04	2

3.1. Helper Node Control Plane

Determining the helper station among the intermittent capable relay stations is very important and also a bottleneck process in the co-operative transference [21]. The cooperative table is formed by the source station and it will select a helper station depending on the data rate. Once the potential helper selection is completed, it will act as a relay node. The data rate ensures the functionality of the CMAC-DR model, which indicates that how effectively the frame is received at the next station. The Helper control flow of the potential relay station initially will be in an idle mode. When the request to send (RTS) frame with data rate information is transferred by the source station and once the CTS frame is transmitted by the destination containing the data rate information, the potential helper overhearing both the frames, CMAC-DRs the RTS frame header extracts and compares the data rate of the source station with its data rate [22]. If it is more than that of the source, it will enter into the cooperation mode and prepares the Helper Ready to Send (HRTS) frame by adding its data rate information in the frame format and broadcasts it to the source station. Similarly, multiple HRTS frames will be sent to the source station from different capable relay stations.

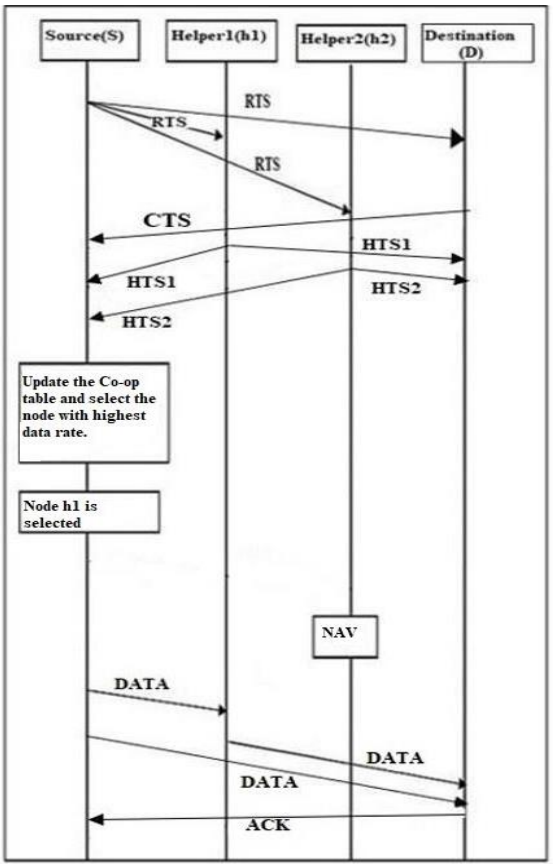


Fig.6. Action flow of CMAC-DR model

3.2. Helper Node Data Plane

Forwarding the data through the selected potential helper station after the exchanging of control frames is a challenging function in the cooperative transference [23]. CMAC-DR model suggests a most appropriate and ideal approach for data forwarding in cooperative transference [24]. The data frame forwarded by the source station is addressed to the selected helper node in which the address of the potential helper is in address field-1 of the data frame. Potential relay/helper station receives the data frame and verifies it whether it is addressed to it, if so, it checks for whether it is in cooperation mode or not. If the potential relay station is in coop mode, it replaces the address of the intended destination in the address field-1 of the data frame and transfers the frame to destination station with its data speed.

3.3. Co-operative Mode Selection

Choosing the transference mode is established by considering three conditions. firstly the intermittent helper nodes checks if it is having any RTS from any of the neighboring node for itself, secondly it is not allocated with Network allocation vector (NAV) from any other station in the network and, thirdly its data rate is more than that of the source station, only if all the above three case are satisfied then cooperative mode will be enabled and broadcasts helper ready to send (HRTS) frame[25].

Algorithm for mode selection by helper nodes

- Initialization.
- Check if it is having any RTS for itself.
- Check if is allocated with NAV from neighboring nodes.
- Check if helper node data rate > source node data rate.
- Transmit helper ready to send (HRTS).
- Else: Be in idle state.

3.4. Actions Performed by Source Node

Initially, source station will be in indolent mode until it has a MPDU to be transferred to the destination [26]. Once the buffer is full and wants to send a frame with body length of l bytes, it first verifies whether the channel is in standby for a DIFS, subsequently the source also waits for its CW to timeout and then it transfers a request to send (RTS) frame in which the data rate information is incorporated to withhold the channel, else it holds back for a random back-off time [27]. If destination side is free for communication it sends a clear to send(CTS) frame with its data rate. If the source station does not obtain a CTS inside “tRTS +tCTS +SIFS”, RTS is re-broadcasted. Next the intermediate potential helper

stations broadcast the helper ready to send (HRTS) frames by keeping their data rate information in the control frame. Now the source station waits for another $t_{Back-off}^{max}$+tHRTS+ SIFS, where $t_{Back-off}^{max}$ is the maximum back-off time for a helper node [20]. If it receives a HRTS within the waiting time, based on the information obtained from the HRTS control frames its cooperation table is updated and the station with maximum data rate is chosen as the helper station and MPDU is transferred through it, else direct transmission is done. The acknowledgment(ACK) has to be obtained within "tACK+16(l+ lh)/2R+2SIFS," where lh denotes length of the header in bytes, else source station would carry out an arbitrary back-off procedure [28]. The l and lh are represented in bytes and data rate(R) with bits per second, thereby the overall transference span for a data frame is given by "8(l+ lh)/2R." Figure.6. indicates the timing diagram of the CMAC-DR model. It considers two intermittent helper stations and a destination. TX and RX represents the transmission and reception line of the respective stations. Here tRTS, tCTS, tHRTS, tACK indicates transmission time for RTS, CTS, HRTS, ACK.

Relay selection algorithm by source node

- Obtaining data rate of the potential helper nodes through HRTS frames sent by them.
- Analyzing different data rate information of all the potential intermittent helper stations to get the relay station with highest data rate.
- If a helper station is chosen, transfer the MPDU through it.
- End

3.5. Enactments of Destination Station

The destination station will be in idle condition until it receives a RTS from the source station. Once the RTS is obtained and if it is self-addressed to itself [29], it will transfer a CTS succeeding a short inter-frame space (SIFS). It will then check for any HRTS frames by any intermediate potential helper stations. If yes, the destination station expects data frame from helper station and also from the source by direct transference. Next if the data frame is successfully obtained from the source station, a positive acknowledgement will be sent to the source indicating a successful transference. Similar optimal relay selection with time slot allocation can be viewed in other works [30].

4. Result Analysis

The stations are located spatially in diverse locale; thus, the fading attribute is also considered for all the stations along with channel fading for the network scenario as shown in Figure.7. The various network attributes that are set for the considered simulation scenario is given by Table.2.

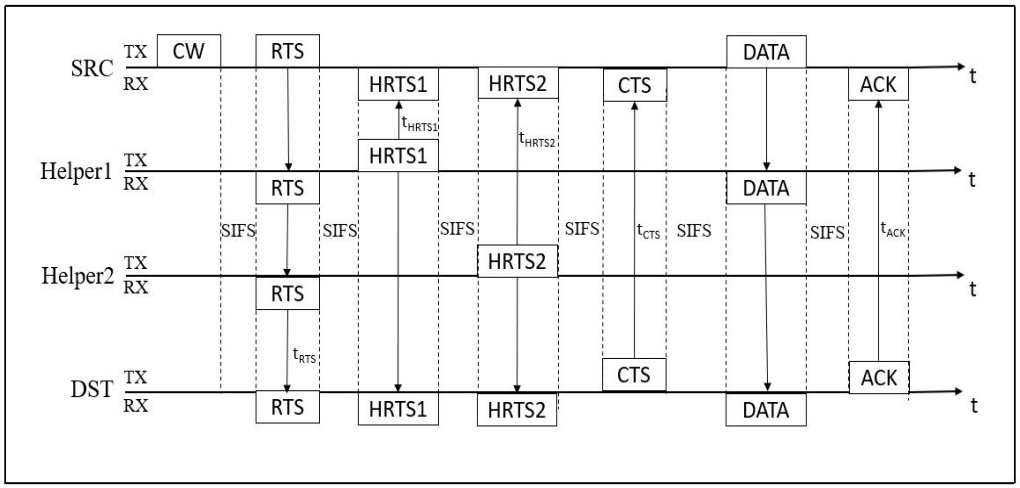


Fig.7. Timing diagram for CMAC-DR model

Simulation results of CMAC-DR model along with legacy 802.11b is evaluated using Qualnet. CMAC-DR model transmits the data frame from source to destination through a helper station based on the data rates. Node number is diversified between 10 to 100 and are located in random in a 600x600 m^2 terrain. The source station transfers packets in the range, 500 to 2313 of size 1024 bytes each with transference rate of 10 packet per second. The simulation time for the scenario considered is 100 seconds. All the stations are enabled with random mobility model. Throughput received, average unicast source to destination delays and total broadcast messages received clearly by the destination station.

Figure.9a. shows that with CMAC-DR model, the throughput given by bits/sec is more as MPDU is transferred cleverly by switching between direct transference and CMAC-DR model.

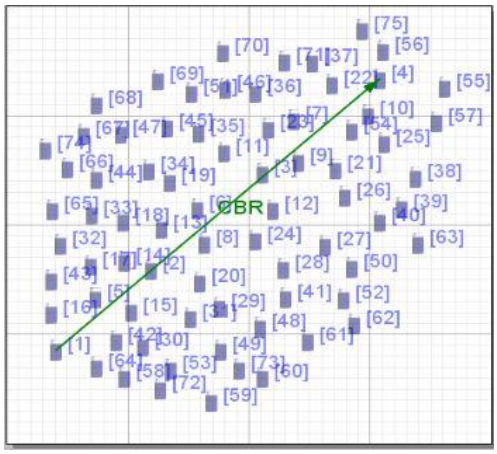


Fig.8. Wireless adhoc system simulation in Qualnet 7.1

Table 2. Simulation parameter

Attributes	Specifications
Node number	10-100
Simulation area (meter sq.)	600x600
Simulation time (seconds)	100
Traffic type	Constant bit rate (CBR)
MAC protocol (wireless)	802.11b
Data rate(MHz)	1, 2, 5.5, 11
Total packets sent	500-2312
Size of each packet	1024

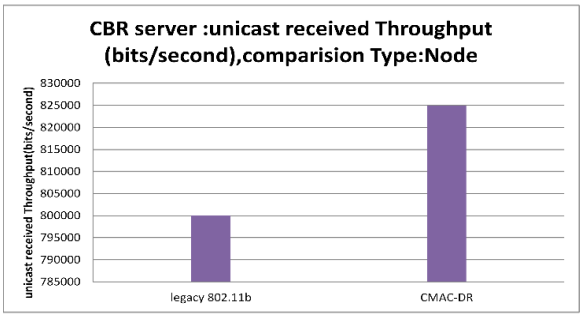


Fig.9a. Throughput of legacy 802.11b vs CMAC-DR model

The received throughput in CMAC-DR model shows increase of 604 metric value in comparison with the IEEE 802.11 protocol for 50 nodes adhoc network. The network lifetime in CMAC-DR model is more, due to the better transference of the MPDU and also because of the availability of stations for long duration.

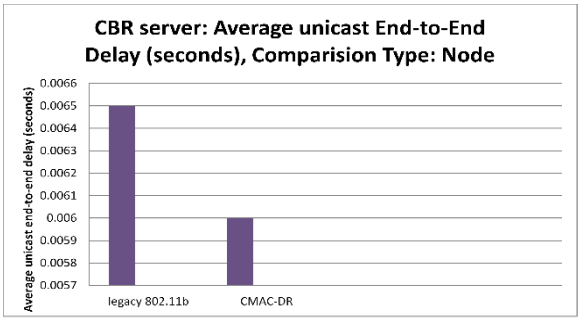


Fig.9b. Delay of legacy 802.11b vs CMAC-DR model

Figure.9b. shows the variation in source to destination delay. In CMAC-DR model frames are transmitted along the best path by choosing best relay node based on the best data rate, as a result delay is reduced in CMAC-DR model. The transmission of frames in legacy 802.11b MAC protocol is only on single-hop communication, whose data rate might be

slow due to channel fading and other parameters, so delay is more.

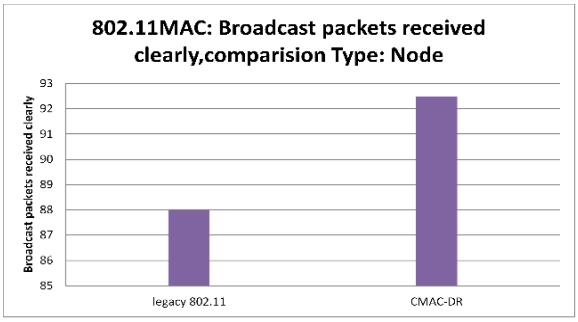


Fig.9c. Broadcast packets received clearly of legacy 802.11b vs CMAC-DR model

Figure.9c. indicates the Broadcast packets received clearly of legacy 802.11b vs CMAC-DR model, Broadcast packets received clearly is more in case of CMAC-DR model because of the employment of both direct transmission mode and also through a potential capable helper station with highest data rate i.e., CMAC-DR model.

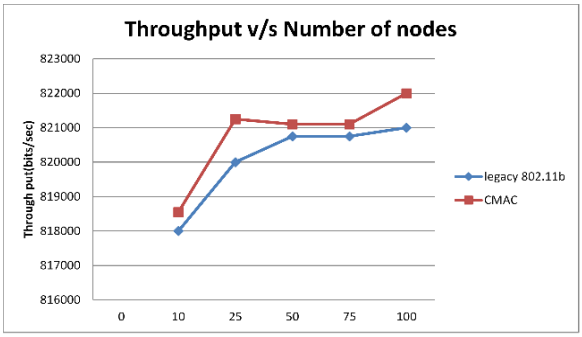


Fig.10a. Contrast in throughput vs number of node

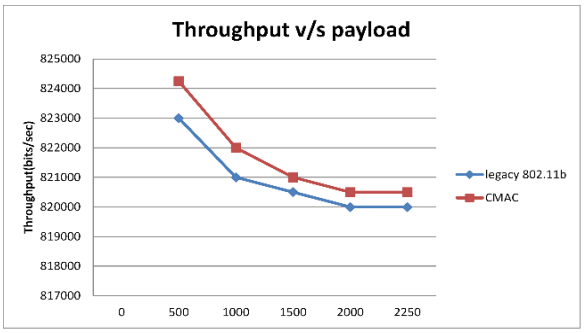


Fig.10b. Contrast in throughput vs payload

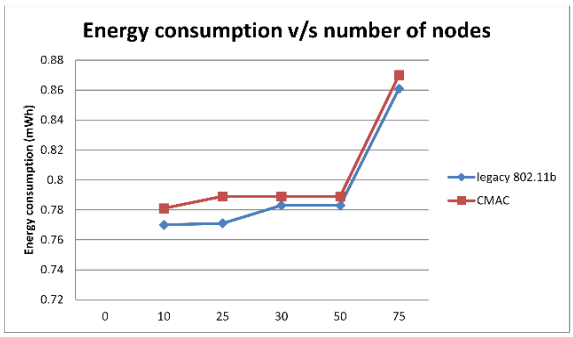


Fig.10c. Contrast in average transmission delay vs number of nodes

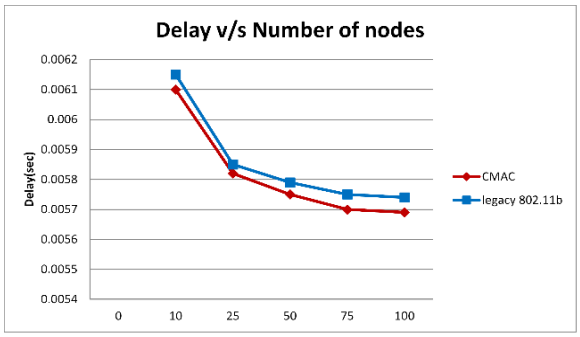


Fig.10d. Contrast in average delay vs number of nodes

The source station transfers 1000 packets to the destination through a wireless channel with various simulation attributes as mentioned in table 2. The simulation is performed for varied set of stations. Figure.10. represents the divergence in QoS, when the of nodes and payloads are varied. Figure. 10(a) & 10(b) indicates the contrast of throughput with node number and payload respectively. Our CMAC-DR model outperforms legacy 802.11b in terms of throughput. The refurbishment is attributed to the inclusion of the data rate as key feature for choosing the capable helper station. The node performance is enhanced, as the source station chooses the intermittent helper station with highest data rate during runtime, which also in turn increases the network lifetime. Average transmission delay provides the information of average latency in successful reception of a data frame. Figure.9c. shows that our CMAC-DR model has less transmission delay than legacy 802.11b, since our model employs a helper station with highest data rate. Also, average delay performance (Figure.10d.) is being enhanced by our CMAC-DR model in comparison with legacy 802.11b protocol. We know that as the number of retransmission limit increases, the delay in successful transmission also increases, we can observe that our CMAC-DR has better delay performance than the legacy 802.11b protocol when the re transmission limit is increased. Our CMAC-DR model is designed and proffered, so as to increase the throughput performance and utilize the channel efficiently for different number of nodes and for variation in payloads with reduced transmission delay. We can infer that CMAC-DR model is finer than IEEE 802.11 protocol by analysing the results obtained from the simulations.

5. Conclusions

The paper proffers CMAC-DR model to enhance the life span of an adhoc network, throughput and delay conduction of the wireless system. The proffered model ensures efficient channel utilization by a station and cleverly selects the capable station with highest data rate for MPDU transference. In a throughput scenario of our CMAC-DR, for a small number of nodes considered, we can see there is 69.2 % increase when compared to Legacy 802.11b and for large number of nodes we can observe 81.2% increase when compared to Legacy 802.11b. When delay of CMAC-DR is compared, for a smaller number of nodes, we can see there is 12.7 % decrease when compared to Legacy 802.11b and for larger number of nodes we can observe 17.3% decrease when compared to Legacy 802.11b.

The CMAC-DR model, precisely selects the mode of data transmission, thus it completely leverages the objective of the cooperative transference i.e., to aid the physical layer to receive the data frame from diverge locale and intelligently extract and combine the data frame so that the quality of transference is enhanced. The aimed objective performance of CMAC-DR model is achieved and it is analysed through numerous simulations and results obtained indicating the enhancement of the system performance. Including different channel fading model, mobility, reducing energy consumption and improving the delay performance is considered as a subsequent study.

References

[1] “Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications”: part 11, IEEE Std for Info tech, 2012.
[2] Thanasis Korakis, etal., “ACooperative MAC (CMAC) Protocol for AdHoc Wireless Networks”, Mitsubishi electric research laboratories”, April 2007.
[3] J.N. Laneman, etal., “Cooperative Diversity in Wireless Networks: Efficient Protocols and Outage Behaviour,” IEEE Tran’s. Info. Theory, Dec. 2004.
[4] J.N. Laneman, etal., “Distributed Space-Time-CMAC-DRd Protocols for Exploiting Cooperative Diversity in Wireless Networks,” IEEE Tran’s. Info theory, Oct. 2003.
[5] T Himsoon, etal., “Lifetime Maximization via Cooperative Nodes and Relay Deployment in Wireless Networks,” IEEE Jrnl on selected areas in Comm’s., 2007.
[6] A. Sendonaris, etal., “User Cooperation Diversity-Part I: System Description,” IEEE Tran’s. Comm’s., Nov. 2003.
[7] S. Cui, etal., “Energy-Efficiency of MIMO and Cooperative MIMO in Sensor Networks,” IEEE Jrnl on selected areas in Comm’s., Aug. 2004.
[8] Xiaofeng Tao, etal., “An overview of cooperative communications”, IEEE Comm’s. Mgz, June 2012.
[9] J Broch, etal., “A Performance Comparision of Multi-hop Wireless Ad Hoc Network Routing Protocols,” Proc. ACM mobile Comm’s., Oct. 1998.

[10] Min Sheng, etal., "Small World Based Cooperative Routing Protocol for Large Scale Wireless Ad Hoc Networks" IEEE Intl Conf. on Comm's., (ICC), 2011
[11] C.E. Perkins, etal., "Ad-hoc On-demand Distance Vector Routing," Proc. IEEE Workshop on mobile computing Sys. and Apps., 1999.
[12] Xiaoyan Wang, etal., "Improving the Network Lifetime of MANETs through Cooperative MAC Protocol Design", IEEE Tran's. On parallel & distributed sys's, April 2015.
[13] A. Ibrahim, etal., "Cooperative Communications with Relay-Selection: When to Cooperate and Whom to Cooperate with?" IEEE Tran's. on wireless Comm's., 2008.
[14] Martin Heusse, etal., "Performance Anomaly of 802.11b", Proceedings of IEEE InfoCom 2003.
[15] "Networking All-in-One Desk Reference for Dummies", by Doug Edition Published by Wiley.
[16] A. Sendonaris, etal., "User Cooperation Diversity-Part I: System Description," IEEE Tran's. Comm's., Nov. 2003.
[17] Xiaofeng Tao, etal., "An overview of cooperative communications", IEEE Comm's. Mgz, June 2012.
[18] J. Wuetal., "Extended Dominating Set and Its Applications in Ad Hoc Networks Using Cooperative Communication,""IEEE Tran's. On parallel &distributed sys's."Aug. 2006.
[19] Weiyan Ge, etal., "Joint Clustering and Optimal Cooperative Routing in Wireless Sensor Networks" IEEE Intl Conf. on Comm's., 2008.
[20] H. Shan, etal., "Cross-Layer Cooperative MAC Protocol in Distributed Wireless Networks," IEEE Tran's. on wireless Comm's., Aug. 2011.
[21] P. Liu,etal., "CoopMAC: A Cooperative MAC for Wireless LANs," IEEE Jrnl on selected areas in Comm's, Feb. 2007.
[22] S Moh, etal., "A Cooperative Diversity-Based Robust MAC Protocol in Wireless Ad Hoc Networks", "IEEE Tran's. On parallel & distributed sys's, 2011.
[23] H. Zhu, etal., "rDCF: A Relay-Enabled Medium Access Control Protocol for Wireless Ad Hoc Networks", IEEE Tran's. On mobile computing, Sept. 2006.
[24] A.K. Sadek, etal., "On the Energy Efficiency of Cooperative Comm's in Wireless Sensor Networks," ACM Tran's on sensor net's., Dec. 2009.
[25] J Zhang, etal., "VC-MAC: A Cooperative MAC Protocol in Vehicular Networks," IEEE Tran's. Vehicular Tech., Mar. 2009.
[26] C. Zhai, etal., "Lifetime Maximization via a New Cooperative MAC Protocol in Wireless Sensor Networks," Proc. IEEE globecom, Dec. 2009.
[27] X. Wang, etal., "NCAC-MAC: Network Coding Aware Cooperateive Medium Access Control for Wireless Networks," Proc. IEEE wireless Comm's. & Networking Conf., Apr. 2012.
[28] S. Kadloor, etal., "Relay Selection and Power Allocation in Cooperative Cellular Networks," IEEE Tran's. on wireless Comm's., May 2010.
[29] S Moh, etal., "A Cooperative Diversity-Based Robust MAC Protocol in Wireless Ad Hoc Networks", "IEEE Tran's. On parallel & distributed sys's, 2011.
[30] E. Beres, etal., "Optimal Relay-Subset Selection and Time-Allocation in DeCMAC-DR-and-Forward Cooperative Networks,"

Authors' Profiles

Nagesh. R. is currently working as Associate Professor Electronics and Communication Engineering Department at Government SKSJ Technological Institute, Bangalore.

He is pursuing his Ph.D. (Engg.) from Visvesvaraya Technological University, Belgaum in the area of Cooperative Computer Networks. He obtained his Master's Degree in VLSI and Embedded Systems from Visvesvaraya Technological University, Belgaum in 2007. He holds Bachelor's degree in Electronics and Communication Engineering from RV College of Engineering, Bangalore in 2002. He has more than 16 years of experience in Teaching and Industry. He was with Wipro Technologies, and TATA Elxsi Ltd. as Senior Chip Design Engineer. His major interest areas of research include Cooperative Computer Networks, MANETS, Vehicular Networks and Embedded Systems.

Dr. M. N. SreeRangaraju received BE Degree in 1985 from Bangalore University and MTech in 1991 from University of Mysore both in Electronics Engineering, and PhD (Telecommunication Engg) from VMU, Salem, TN, India in 2011.Professor of Electronics and Communication Engineering at Bangalore Institute of Technology (BIT), Bengaluru, India has been in Teaching UG and PG Students of Electronics and Telecommunication Engineering students for nearly Thirty Years. He has organized several National level conferences and workshops in this tenure. He has published his Research papers in Prestigious IEEE Conferences held in USA, China, Malaysia and Egypt. He also has served has Session Chair for several National and International Conferences held in India and aboard. He also has served as editor of several journals and Magazine. He holds life member of ISTE, MVLSI, and IMAPS and also member of IEEE. His research interests include mobile and wireless communications and networks, personnel communication services and high-speed communication routing protocols and wireless channel modeling.

N. R. Kushar Gowda completed his Bachelor's degree in Electronics and communication Engineering from Government SKSJ Technological Institute, Bengaluru in 2020. He has worked as a research intern at Nihon Communications limited, Bengaluru on a project to establish a cooperative scheme at MAC sublayer in wireless Adhoc networks. He has also presented a technical research paper at 3rd International conference on advanced production and Industrial engineering organized by Delhi technological university. He has also participated in Global hack the crisis hackathon to combat Novel corona widespread by designing a "Portable and Affordable Ventilator with Assist control for Novel corona virus- PAVAN" and secured second place in crisis response track of the hackathon. The project was one among the best top-30 ideas in Hack the crisis, India hackathon organized by Ministry of Electronics and Information Technology, Government of India. His major areas of interest include wireless communication and networks, Artificial Intelligence and Machine Learning.

Vinayak Shekharappa Antin completed his Bachelor's degree in Electronics and communication Engineering from Government SKSJ Technological Institute, Bengaluru in 2016. He Has worked with Volvo group as Lead ESW application engineer with skills AUTOSAR, Model Based, Agile, Embedded Linux. He was senior software engineer at HCL Technologies working on Configuration of AUTOSAR RTE (Design of Software Components (SWC's) and Integrate SWC's into the ECU architecture (Service, Data and Task mapping). Functional Development with Embedded C and MATLAB (SIMULINK, State flow). Writing Software Requirement specification for End to End functionality. He is presently working as a project engineer at FEV Creltech GmbH, Germany.

How to cite this paper: Nagesh R., M. N. Sree Rangaraju, N. R. Kushar Gowda, Vinayak Shekharappa Antin, "Cooperative MAC Protocol based on Best Data Rate (CMAC-DR)", International Journal of Computer Network and Information Security(IJCNIS), Vol.16, No.3, pp.26-36, 2024. DOI:10.5815/ijcnis.2024.03.03

Int J Syst Assur Eng Manag
https://doi.org/10.1007/s13198-024-02380-1

REVIEW PAPERS

Deep learning in neglected vector-borne diseases: a systematic review

Atmika Mishra[1] · **Arya Pandey**[1] · **Ruchika Malhotra**[1]

Received: 5 April 2024 / Revised: 20 May 2024 / Accepted: 23 May 2024
© The Author(s) under exclusive licence to The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2024

Abstract This study explores the application of Deep Learning in combating neglected vector-borne Diseases, a significant global health concern, particularly in resource-limited areas. It examines areas where Deep Learning has proven effective, compares popular Deep Learning techniques, focuses on interdisciplinary approaches with translational impact, and finds untapped potential for deep learning application. Thorough searches across multiple databases yielded 64 pertinent studies, from which 16 were selected based on inclusion criteria and quality assessment. Deep Learning applications in disease transmission risk prediction, vector detection, parasite classification, and treatment procedure optimization were investigated and focused on diseases such as Schistosomiasis, Chagas disease, Leishmaniasis, Echinococcosis, and Trachoma. Convolutional neural networks, artificial neural networks, multilayer perceptrons, and AutoML algorithms surpassed traditional methods for disease prediction, species identification, and diagnosis. The interdisciplinary integration of Deep Learning with public health, entomology, and epidemiology provides prospects for improved disease control and understanding. Deep Learning models automate disease surveillance, simplify epidemiological data processing, and enable early detection, particularly in resource-constrained settings. Smartphone apps driven by deep learning allow for rapid disease diagnosis and identification, boosting healthcare accessibility and global health outcomes. Improved algorithms, broadening the scope of applications to areas such as one health approach, and community engagement, and expanding deep learning applications to diseases such as lymphatic filariasis, hydatidosis, and onchocerciasis hold promise for improving global health outcomes.

Keywords Deep learning · Neglected diseases · Vector-borne diseases · Neural network

JEL Classification O33 · I10

1 Introduction

In the field of medical diagnostics, Deep Learning (DL) techniques have become increasingly potent instruments with the potential to transform decision-making processes completely. DL models have been applied to predict transmission risks and outcomes of neglected Vector-Borne Diseases (VBDs) in recent years. VBDs continue to be a major global health challenge, especially in settings with limited resources.

For example, studies on schistosomiasis mansoni and schistosomiasis japonica have demonstrated the significance of variables like water contact, geographic variances, and the history of infection in predicting the spread of the disease (Xu et al. 2013; Hammad et al. 1996). These studies have demonstrated that DL algorithms can predict disease outcomes with high efficacy and accuracy, sometimes even outperforming conventional logistic regression models. Similarly, research on cutaneous leishmaniasis by Bamorovat et al. (2021) has shown that DL has the potential to predict disease outcomes by using machine learning (ML) techniques to predict cases that will not respond.

Additionally, DL has been used to identify Chagas disease vectors automatically, with high accuracy rates that can

✉ Atmika Mishra
atmikamishra_se20b8_19@dtu.ac.in

[1] Department of Software Engineering, Delhi Technological University, Main Bawana Road, Delhi, NCT of Delhi 110042, India

Published online: 10 June 2024

greatly enhance surveillance and control operations (Khalighifar et al. 2019; Parsons and Banitaan 2021). Strong correlations between predicted and true values have also been observed when using DL models to predict the Leishmania infantum load in dogs, which may help with disease surveillance and control (Torrecilha et al. 2017).

Other studies have demonstrated high accuracy rates in classifying complex image datasets related to VBDs, with a focus on the identification of snails and parasites of medical importance for schistosomiasis (Tallam et al. 2021; Liu et al. 2019). To improve diagnostic efficiency and accuracy, DL models have also been used to automatically classify various hepatic cystic echinococcosis subtypes from ultrasound images by Wu et al. (2022).

Furthermore, DL has been used to model optimization of benznidazole chitosan microparticles, demonstrating its potential for drug formulation optimization and key property prediction for disease treatment by Leonardi et al. (2009). Similarly, supervised artificial neural networks (ANNs) has been used to predict the pharmacological properties of compounds in drug discovery and treatment prediction for tropical diseases such as Chagas disease by Bamorovat et al. (2021).

This organized methodology enabled a thorough examination of DL's impact on neglected VBD research and its possible implications for global health outcomes.

1.1 Organisation of the study

The study begins with Sect. 1, which introduces the topic of this study. Section 2 contains findings of related published work, and explores the popular applications of DL in neglected VBDs. Section 3 represents Material and Methods explaining the article selection process and PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework for this study, including search strategy, inclusion criteria, quality assessment (QA) and final selection of the studies. The results and discussion Sect. 4 summarises the findings from the selected studies for the RQs of this study. Finally, Sect. 5 concludes the review by summarising the findings of the study and highlighting important avenues for future research.

2 Related work

DL techniques have been applied in several neglected VBDs to enhance disease diagnosis, control, and prediction. For example, studies on the application of DL for identifying and categorising parasites and vectors have been carried out. DL models have been used to identify snails and parasites of medical significance, which aids in the early diagnosis and treatment of schistosomiasis. Similarly, by categorizing and segmenting parasites, DL has been utilized to increase the accuracy of Chagas disease diagnosis.

Furthermore, DL has been used to create risk assessment tools and models to predict disease propagation. For example, research has employed DL algorithms to examine environmental variables that influence the spread of visceral leishmaniasis, offering valuable insights into the disease's dynamics and guiding focused intervention approaches. Additionally, DL has been applied to analyze satellite imagery and geographical data to predict the prevalence and distribution of VBDs, such as malaria and dengue, aiding in the development of early warning systems and effective disease control measures.

Moreover, DL has been applied to improve monitoring and surveillance programs for diseases. Large-scale epidemiological data can be analyzed using DL models to find patterns and trends in the spread of disease. This information helps health authorities allocate resources more wisely and launch interventions on time. These studies show how DL methods can help us better understand and treat neglected vector-borne illnesses, opening new avenues for enhanced disease prevention and control tactics.

Recent surveys have shown that ML and DL have substantial potential applications in diagnosing and treating VBDs. Ghosh and Dasgupta (2022) address the economic and health consequences of neglected tropical diseases, including VBDs, and highlight how ML techniques are being used to screen possible therapeutic compounds, assisting in the hunt for new drug targets. This paper emphasizes the importance of novel ML applications in solving the ongoing difficulties of neglected tropical diseases. Similarly, Yu et al. (2023) present an overview of common DL algorithms, such as ANNs, Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN), focused on their use in disease prediction. They highlight DL's exceptional accuracy in medical applications, which frequently outperforms existing approaches, supporting its usage in disease prediction and diagnosis, including VBDs.

Further exploring the area, Sharma et al. (2022) investigate several DL architectures and their considerable impact on medical diagnoses. Their work emphasizes the benefits of DL, such as increased accuracy and efficiency, while also emphasizing the hurdles that must be overcome before widespread implementation. Shaikh et al. (2023) investigate different supervised ML algorithms for detecting diseases such as malaria, chikungunya, and dengue fever, with a specific focus on VBDs. They discovered that the modified Support Vector Machine (mSVM) had the highest accuracy, suggesting the promise of ML techniques to improve VBD diagnosis and management. Kaur et al. (2021) also critically analyze ML and DL methods for identifying and forecasting fatal disease outbreaks to limit the risk of disease spread. Collectively, these surveys demonstrate the impact of ML

and DL in the medical sciences, notably in improving the diagnosis and treatment of VBDs, and have the potential to be applied to neglected VBDs.

2.1 Research gaps and motivation for the study

In the current literature, an important gap in research addressed by our review is the lack of holistic analysis in DL approaches applied to neglected VBDs. Recent research tends to focus on individual applications rather than broad appraisals of DL's capabilities across all facets of disease management and control. Furthermore, primary VBDs, such as Dengue and Malaria, receive far more attention than other neglected VBDs, which are comprehensively discussed in this review.

While existing research focuses mostly on ANNs and CNNs, it overlooks some important topics and techniques within the DL paradigm. For example, the use of techniques such as Gated Recurrent Units (GRUs) for sequential analysis of data, generative adversarial networks (GANs) for synthetic data generation, and Reinforcement Learning (RL) for tailored treatment planning remains unexplored. Additionally, Graph Neural Networks (GNNs) and Explainable AI (XAI) approaches are notably absent from current research discourse. By highlighting these research gaps, this review provides a more comprehensive knowledge of the potential of DL in managing neglected VBDs.

Studies may demonstrate the efficacy of DL models in specific circumstances, but there is at times a dearth of discussion about the obstacles encountered and potential areas for development. This study fills these gaps by critically examining methodology, identifying strengths and limits, and recommending future research areas to improve the efficacy of DL strategies in neglected VBD control and management.

3 Material and methods

Following a comprehensive analysis of 64 papers from a variety of databases, including IEEE Xplore, Google Scholar, and PubMed, 16 studies that used DL to diagnose neglected VBDs, namely, Cutaneous Leishmaniasis, Hepatic Echinococcosis, Trachoma, Schistosomiasis, and Chagas Disease, were chosen for inclusion. These studies met the requirements for inclusion such as relevance, research thoroughness, peer review status, English language publication, and a focus on neglected VBDs.

3.1 Overview

Figure 1 contains the PRISMA flow diagram of this study, explaining the process by which 64 publications about neglected VBDs were found after a thorough search of IEEE Xplore, Google Scholar, and PubMed. In the first instance, we found 11 duplicates in the dataset, which were eliminated. Thereafter, we carefully went through 53 studies and excluded 16 that did not fit our predetermined inclusion requirements.

After the screening procedure, we identified 37 studies for retrieval. After retrieval, we thoroughly evaluated these studies to determine their eligibility and eliminated 21 studies for failing to meet the quality standards. 16 papers that satisfied all requirements made it to the final selection and were reviewed.

3.2 Research questions

Examining the current state of DL applications in neglected VBD research is the goal of this review. Of special interest is how these applications impact disease diagnosis, prognosis, and treatment. Additionally, the study attempts to comprehend the underlying reasons for the adoption of DL approaches for different VBDs in different geographical regions. According to our research goal, the following RQs have been developed:

(1) RQ1: What are the distinct sub-areas within neglected VBDs where DL techniques have been applied?
(2) RQ2: What is the scope of DL applications within each neglected VBD, and how do different DL models perform in disease-specific analysis?
(3) RQ3: How can interdisciplinary integration of DL with fields such as epidemiology, entomology, and public health improve understanding and control of neglected VBDs?
(4) RQ4: What emerging areas within neglected VBDs research present potential for DL applications, and how might leveraging DL in these areas lead to innovative solutions and advancements in global health outcomes?

3.3 Article selection

We intended to ensure a comprehensive and diverse selection of relevant literature about the study's topic by utilizing referred databases. An iterative and systematic process was employed to generate the search phrases. After creating a list of terms, the process was repeated until no more significant publications could be located. The studies published from July 1996 to November 2023 were considered for our review.

Logical operators such as AND and OR were used to create links between search terms and their synonyms. OR was used inside a group of synonyms and AND was used across groups to lessen the possibility of missing relevant research. To ensure thoroughness, references from the

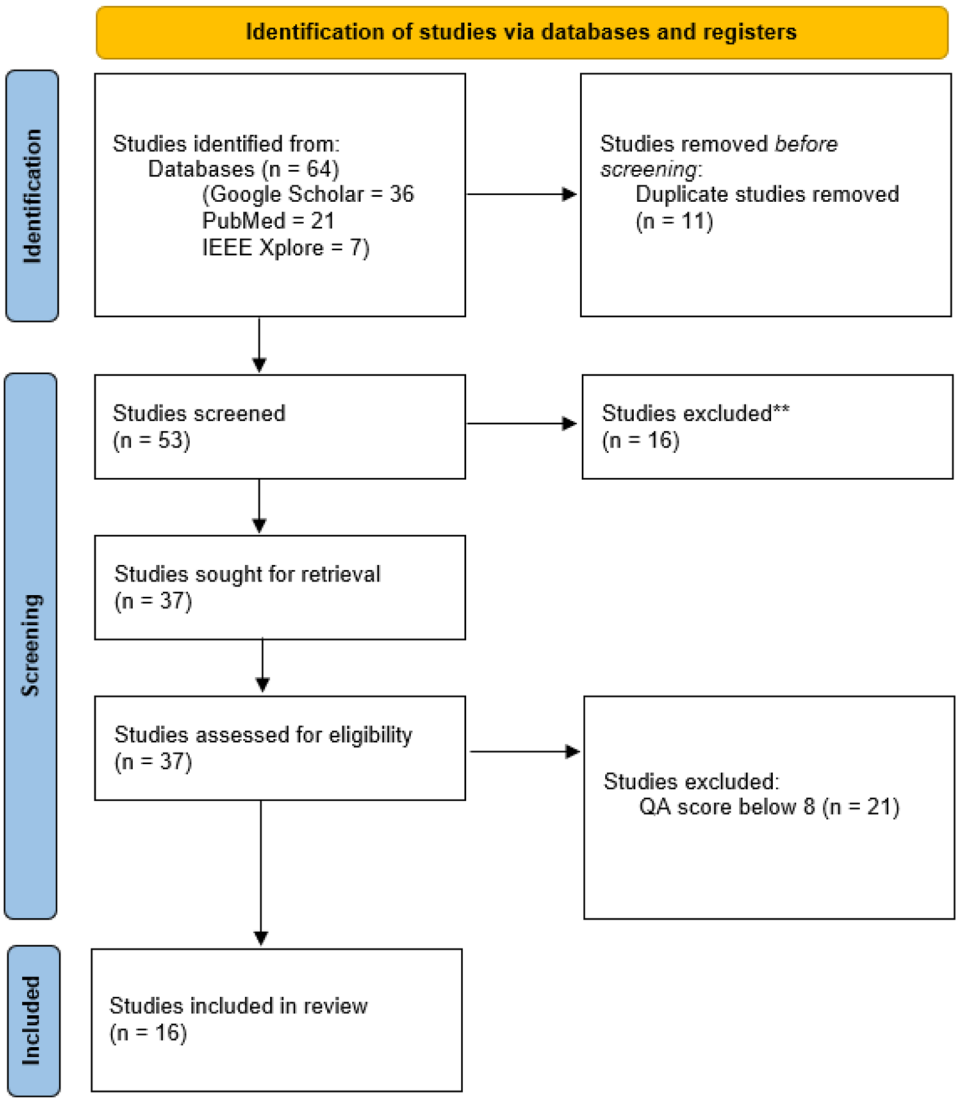


Fig. 1 PRISMA Flow Diagram of the Review. ** - did not meet the inclusion criteria, QA - Quality Assessment

collected research were manually cross-checked before being included.

Table 1 contains the search strategy for the study which combines the key concepts of our search question which in turn helps us to retrieve the required results.

3.4 Inclusion and exclusion criteria

When conducting this literature review, the following criteria were considered for inclusion:

(1) Studies that were published between July 1996 and November 2023.
(2) Studies that are cited by other researchers.
(3) Studies that are relevant to the application of DL in neglected VBDs.

Included studies were restricted to those published in English with citations and that were electronically accessible. If a conference study had a more recent journal version, only those published later were included. The suitability of each study for use in the diagnosis of tropical diseases was carefully assessed. Every search started with the title and keywords.

Table 1 Search strategy

("Trachoma" OR "Fascioliasis" OR "Leishmaniasis" OR "Lymphatic Filariasis" OR "Schistosomiasis" OR "Hydatidosis" OR "Echinococcosis" OR "Chagas" OR "Taeniasis" OR "Cysticercosis" OR "Onchocerciasis" OR "Mycetoma") AND ("Deep Learning" OR "Neural Network" OR "Neural Networks")

We also reviewed the introduction for publications where the information could not be inferred from the relevant sections. Following analysis, the entire text was downloaded for further reading and data extraction if these items satisfied our requirements.

During the literature review phase of this investigation, the following exclusion criteria were taken into account:

(1) Studies that are not addressing the RQs.
(2) Studies whose manuscript is not available in English.
(3) Studies on the neglected VBDs, where the focus is not on DL.

3.5 Quality assessment

Investigators independently assessed the quality of the included studies, which were all discussed, based on the QA of Diagnostic Accuracy Studies. The risk of bias was also studied in which data source, sample size, coherency of results, etc were analyzed. There were 10 criteria for QA of the collected studies based on applicability to our research problem, clarity in objectives and methodologies, availability of data information, declaration of comprehensive results and conclusions, and addressing limitations. These 10 criteria are comprehensively stated in Table 2 under the QA Questionnaire. The questions were rated: Yes (Y): 1 score, Partial (P): 0.5 score, and No (N): 0 score. Upon completion of the assessment, studies with a total score of 8 or above were selected. Table 3 showcases the QA scores calculated for the selected studies.

3.6 Final articles

The final selection included 16 papers that demonstrated the potential of DL and ML to address issues in neglected VBDs' control and management by emphasizing their roles in disease prediction, diagnostics, and treatment strategies.

4 Results and discussion

4.1 Research question 1

DL models have been investigated in many studies to predict the risks of schistosomiasis mansoni and japonica transmission (Xu et al. 2013; Hammad et al. 1996). In predicting the spread of disease, these studies highlight the importance of factors like prior infection history, contact with water, and regional variations. Furthermore, the application of DL is useful to forecast the course of diseases, as compared to conventional logistic regression models. ML techniques were applied in the cutaneous leishmaniasis study by Bamorovat et al. (2021) to predict unresponsive cases, demonstrating the potential of ML in disease outcome prediction.

The use of DL methods for the automated identification of Chagas disease vectors has been the focus of research by Khalighifar et al. (2019) and Parsons and Banitaan (2021). These studies show that identifying vector species with high accuracy rates can greatly enhance surveillance and control efforts. More investigation to integrate multimodal data sources, including genetic, clinical, and environmental data, could lead to comprehensive diagnostic tools that account for the complex interplay between factors that influence the course of a disease and the efficacy of a treatment strategy.

DL models have been used in studies to identify medically significant parasites and snails that cause schistosomiasis by Tallam et al. (2021) and Liu et al. (2019). These studies show that complex image datasets linked to VBDs can be classified with high accuracy rates, sometimes even surpassing the performance of highly skilled human parasitologists. DL models were used in the hepatic cystic echinococcosis

Table 2 Quality assessment questionnaire

QNo.	Quality assessment question
QA1	Are the objectives of the research clearly stated?
QA2	Is the research pertinent to the illnesses included in this review?
QA3	Does the study clarify the source of data for the chosen VBDs analysis?
QA4	Are the techniques for gathering data adequately described and supported by the goals of the study?
QA5	Does the study design and methodology provide a clear description of the DL techniques used in the investigation?
QA6	Are the conclusions and results expressed clearly?
QA7	Do the findings corroborate the conclusions?
QA8	Are the study's limitations stated clearly?
QA9	Does the study add value to the body of literature?
QA10	Does the study answer the research questions of the review?

QA Quality assessment

Table 3 Quality assessment scores for selected studies

Paper	QA1	QA2	QA3	QA4	QA5	QA6	QA7	QA8	QA9	QA10	Score
Xu et al. (2013)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	10
Hammad et al. (1996)	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	10
Tallam et al. (2021)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	10
Liu et al. (2019)	Y	Y	Y	Y	Y	Y	P	N	P	Y	8
Khalighifar et al. (2019)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	10
Parsons and Banitaan (2021)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	10
Guerra et al. (2013)	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	9
Leonardi et al. (2009)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	10
Bamorovat et al. (2021)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	10
Torrecilha et al. (2017)	Y	Y	Y	Y	Y	Y	P	Y	P	Y	8.5
Zheng et al. (2018)	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	9
Wu et al. (2022)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	10
Xin et al. (2020)	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	9
Yang et al. (2023)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	10
Wang et al. (2023)	Y	Y	Y	Y	Y	Y	P	Y	P	Y	9
Milad et al. (2023)	Y	Y	Y	P	Y	Y	Y	Y	Y	Y	9

Y Yes, *N* No, *P* Partial, *QA* Quality assessment

study to automatically classify various subtypes from ultrasound images, improving the efficiency and accuracy of the disease's diagnosis (Wu et al. 2022). Research has employed ANNs for chitosan microparticles by Leonardi et al. (2009), showcasing the potential of DL in optimizing drug formulations and predicting key properties for treating Chagas' disease. To predict the pharmacological characteristics of compounds, the Chagas disease study by Guerra et al. (2013) used supervised ANN. This highlights the potential of DL in the process of finding new drugs and treating VBDs.

Using a Radial Basis ANN, the study on Leishmania infantum load prediction in dogs by Torrecilha et al. (2017) demonstrated a strong correlation between predicted and true values. In areas with limited resources, this may be helpful for Leishmania infantum surveillance. The success of DL for diagnostic purposes, especially in automating the classification of disease-related images and predicting treatment outcomes, like the development of smartphone applications for diagnosing trachoma and the classification of ultrasound pictures for hepatic cystic echinococcosis, shows how technology can democratize access to healthcare. Enhancing the interpretability and applicability of DL models should be the main goal of research, particularly in situations where data availability is inconsistent or limited.

Table 4 provides a comprehensive overview of the performance metrics for various models used in detecting Schistosomiasis Japonica, Schistosoma mansoni infection rates, and Schistosomiasis in the Senegal River Basin, highlighting their effectiveness and aiding in selecting the most suitable model for diagnosis.

Table 5 gives an overview of the performance metrics of Chagas disease vectors. The table helps depict the superiority of a few models in identifying vectors of different species.

Table 6 outlines the performance metrics for another neglected VBD, Leishmaniasis, where the Multilayer Perceptron (MLP) and Radial Basis ANN models show notable performance with a sensitivity of 90.3% and an area under the curve (AUC) of 0.88. These values are indicative of their reliability in diagnosing the leishmaniasis disease.

Table 7 details the performance metrics for echinococcosis disease, highlighting the high AUC values of the Deep Convolutional Neural Networks (DCNN) models (ResNet-50, ResNext-50, and VGG-16 (Visual Geometry Group 16-layer model)) for cystic (0.974) and alveolar echinococcosis (0.981), underscoring their effectiveness in detecting these conditions.

Table 8 summarizes performance metrics for Trachoma. The Automated Machine Learning (AutoML) models achieved impressive Area under the Precision-Recall curve (AuPRC) values. The AuPRC values of 0.945 and 0.975 for trachomatous inflammation-follicular (TF) and trachomatous inflammation-intense (TI) models, respectively were found.

This showcases their precision in diagnosing trachomatous inflammation. These models can also be applied to other VBDs where current research is scarce to improve disease control.

The boxplot of model accuracies for various Schistosomiasis models in Fig. 2 highlights the exceptional performance of the CNNs used for snail classification in the Senegal River Basin. This model, showing an accuracy of 99.64%, marks it as an outlier, with very high success in snail classification. This outlier status is evident, as it significantly outperforms

Table 4 Performance metrics for schistosomiasis japonica, schistosoma mansoni infection rates, and schistosomiasis in the senegal river basin

Disease	Model	Performance metric
Schistosomiasis Japonica	NN	Sensitivity 83% Positive predictive value 63%
	VGG16	Accuracy 99.60%
	InceptionResNetV2	Accuracy 91.21%
	CNNs	Snail classification 99.64% Parasite classification 88.14%
Schistosoma mansoni infection rates	NN	Sensitivity 83% Positive predictive value 63%
Schistosomiasis in the Senegal River Basin	CNNs	Snail classification 99.64% Parasite classification 88.14%

NN Neural network, *CNN* Convolutional neural network, *VGG16* Visual geometry group 16-layer model

Table 5 Performance metrics for chagas disease (Vectors)

Disease	Model	Performance metric
Chagas disease vectors	TensorFlow with DNNs	Correct identification rate Mexican species 83.0–95.8% Brazilian species 86.7–98.9%
	PCA + DT, PCA + RF	Accuracy: 100%
	Supervised ANNs	Training performance 84% External validation Performance: 77%

PCA Principal component analysis, *DT* Decision tree, *RF* Random forest, *DNN* Deep neural network

Table 6 Performance metrics for Leishmaniasis

Disease	Model	Performance metric
Leishmaniasis	MLP	Overall accuracy 87.8%
	MLP, Radial Basis ANN	Sensitivity: 90.3% Specificity: 86% AUC: 0.88

MLP Multilayer perceptron, *ANN* Artificial neural network, *AUC* Area under the curve

other accuracies of the other models whose accuracy can be found between 83% and 99.6%.

Additionally, the high mean accuracy shown in the boxplot is indicative of the fact that the nominal models are reliable and potentially applicable to other diseases. This also suggests that the robust modeling approach, particularly for the CNN models, is effective and can be adapted for the detection and classification tasks of other neglected VBDs.

Similarly, the boxplot of model accuracies for Chagas disease in Fig. 3 reveals the outstanding performance of Principal Component Analysis (PCA) combined with

Table 7 Performance metrics for echinococcosis

Disease	Model	Performance metric
Echinococcosis	airPLS-KS-BP	True positive rate: 94.2857 ± 4.0721%
		True negative rate: 95.2381 ± 0%
		Overall accuracy rate: 94.6939 ± 2.3269%
	VGG19, Inception-v3, ResNet18	Pretrained accuracy: 88.2–90.6%
		Scratch accuracy: 69.4–75.1%
	ResNet-50, ResNext-50, VGG-16	AUC: 0.913–0.986
	DCNN (ResNet-50, ResNext-50, VGG-16)	Cystic Echinococcosis AUC: 0.974
		Alveolar Echinococcosis AUC: 0.981

airPLS Adaptive iteratively reweighted penalized least squares, *KS* Kennard-Stone, *BP* Back propagation, *DCNN* Deep convolutional neural network, *AUC* Area under the curve, *VGG16* Visual geometry group 16-layer model, *VGG19* Visual geometry group 19-layer model

Table 8 Performance metrics for trachoma

Disease	Model	Performance metric
Trachoma	AutoML (TF model)	AuPRC: 0.945
	AutoML (TI model)	AuPRC: 0.975

AuPRC - Area Under the Precision-Recall Curve, AutoML - Automated Machine Learning, TF - Trachomatous Inflammation-Follicular, TI - Trachomatous Inflammation-Intense

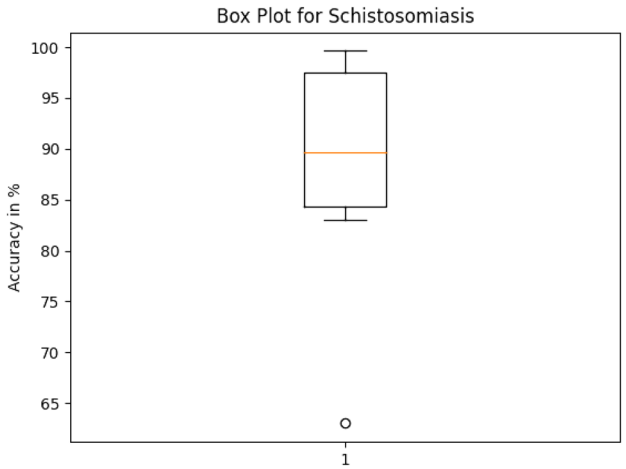


Fig. 2 Box Plot of model accuracy for Schistosomiasis

Decision Tree (DT) and Random Forest (RF), which achieved a perfect accuracy of 100%, making it a clear outlier.

This is notable as it surpasses other models, which have identification rates between 83.0% and 98.9% for the TensorFlow + Deep Neural Network (DNN) model in Mexican and Brazilian species of the Chagas disease vector. The high mean accuracy displayed in the boxplot further suggests that these models are reliable and versatile enough to be applied to other diseases.

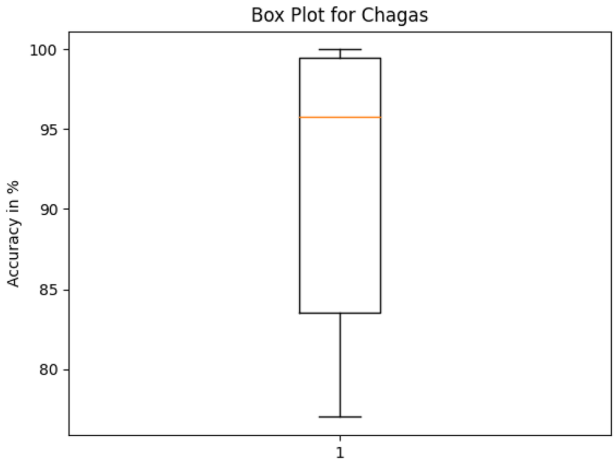


Fig. 3 Box Plot of model accuracy for Chagas

This highlights the robustness of the PCA + DT and PCA + RF models, indicating their potential effectiveness for a wider range of disease detection and classification.

Research has employed supervised ANNs to forecast the pharmacological characteristics of substances related to neglected VBDs such as in the study by Guerra et al. (2013). In the trachoma study, Milad et al. (2023) demonstrated how image processing applications can democratize access to artificial intelligence in healthcare by creating an AutoML smartphone application for trachoma diagnosis.

The research studies highlight a comprehensive exploration of the application of DL techniques in addressing various challenges related to VBDs. One notable aspect is the utilization of these advanced algorithms in disease prediction, where factors such as prior infection history, environmental conditions, and regional variations play crucial roles.

These studies underscore the importance of leveraging ML for accurate forecasting while demonstrating its superiority over traditional logistic regression models in most cases. Moreover, the automated identification of disease

vectors and medically significant parasites showcases the potential of DL in enhancing surveillance and control efforts, particularly in resource-constrained regions. However, while these studies provide valuable insights, further research could delve deeper into developing more robust models that consider additional variables such as climate change patterns, human mobility, housing conditions, and socioeconomic factors, which can significantly influence the dynamics of VBDs. Additionally, exploring interdisciplinary approaches by integrating epidemiological data with satellite imagery and mobile health technologies could offer new avenues for real-time monitoring and early warning systems.

To ensure fair access and patient safety in the age of artificial intelligence-driven diagnostics and treatment algorithms, it is also imperative to investigate the ethical implications and potential biases associated with algorithmic decision-making in healthcare settings.

4.2 Research question 2

Applications of DL have demonstrated a great deal of promise in treating neglected VBDs, with each illness offering its own set of opportunities and challenges. DL models, such as CNNs and back-propagation artificial neural networks (BP ANN), have been used for classification and predictive modeling in the case of schistosomiasis. BP ANN, for example, performed well in predicting transmission risks, as evidenced by significant metrics such as correlation coefficient and mean squared error (MSE) as depicted by Xu et al. (2013). Furthermore, CNNs demonstrated exceptional accuracy in the classification of schistosomiasis-related snails and parasites, with classification accuracy for snails reaching as high as 99.64% and for parasites as high as 88.14% by Liu et al. (2019).

DL applications have also helped with Chagas disease, especially with automating the identification of disease vectors. With accurate identification rates of over 80% for both Mexican and Brazilian species of Chagas disease vectors, DL algorithms greatly increased identification rates (Khalighifar et al. 2019). Furthermore, high accuracy rates in predicting Anti-Trypanosoma cruzi activity were obtained through the use of supervised ANNs, which is an important aspect of drug development for the treatment of Chagas disease (Guerra et al. 2013).

DL techniques have been used for prognostic and diagnostic purposes in leishmaniasis studies. For example, an MLP classifier with an AUC of 0.88 and an overall accuracy of 87.8% was able to predict unresponsive cases of cutaneous leishmaniasis by Bamorovat et al. (2021). Similarly, DL models such as radial basis ANNs have demonstrated potential in anticipating parasite load, which can help with disease management in Torrecilha et al. (2017).

Through DL applications, echinococcosis-another neglected VBD-has seen improvements in diagnosis and classification. DCNNs are among the models that have shown to perform better in identifying lesions associated with hepatic echinococcosis from imaging data; segmentation of these lesions achieved a Dice score of 89.89% by Xin et al. (2020). Moreover, DL models achieved accuracies ranging from 88.2 to 90.6% in the classification of hepatic cystic echinococcosis, outperforming conventional methods (Wu et al. 2022).

Novel DL applications for diagnosing Trachoma, an under-researched tropical disease affecting the eyes, have also been developed. High diagnostic properties were demonstrated by AutoML algorithms, which were able to identify TF and TI from conjunctival images with respective AuPRC values of 0.945 and 0.975 (Milad et al. 2023).

The analysis highlights how DL has the potential to revolutionize the treatment of neglected vector-borne illnesses such as trachoma, echinococcosis, leishmaniasis, schistosomiasis, and Chagas disease. Predictive modeling, species identification, and disease diagnosis have all shown impressive levels of accuracy when using DL models, such as CNNs and MLP classifiers.

More investigations are necessary to determine whether DL solutions can be scaled and applied to varied geographical locations and demographic groups. To take into account the complex nature of illness management and transmission, future research should concentrate on combining various data sources and creating hybrid DL models.

Research gaps still exist despite the advancements, such as the requirement for long-term studies to evaluate the effectiveness of DL-based therapies over the long run and initiatives to improve model interpretability.

Figure 4's histogram is essential as it compares the best models from various neglected VBDs, offering a concise overview for selecting effective models in our study. To summarize, DL applications have shown great promise for multiple neglected VBDs, encompassing diagnosis, prognosis, and predictive modeling and classification. In disease-specific analysis, models such as CNNs and MLP classifiers have demonstrated impressive accuracy, supporting risk assessment, species identification, and prognosis. More research is necessary to investigate the scalability and generalizability of DL models across various geographic locations and demographic groups impacted by these diseases.

Continuous research is necessary to modify DL approaches to new problems brought forth by urbanization and climate change. Research projects that include many disciplines can work together to create complete DL frameworks customized to the particular epidemiological conditions of neglected VBDs. This will help advance global disease surveillance, diagnosis, and control programs.

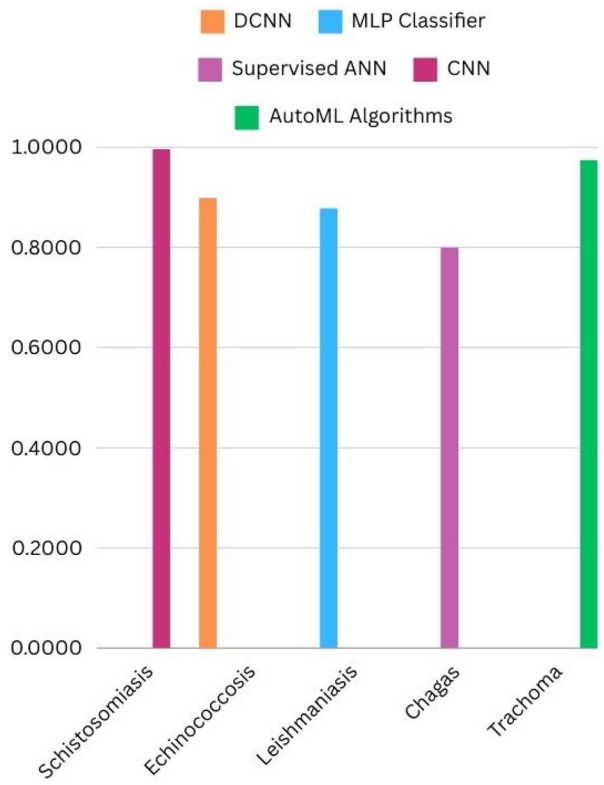


Fig. 4 Comparison of the accuracy of the best models of each neglected VBDs

4.3 Research question 3

DL's interdisciplinary integration with public health, entomology, and epidemiology offers a huge chance to improve our knowledge and ability to control neglected vector-borne illnesses. Advanced computer methods, such as DL models, can be used to understand the intricate dynamics of vector behavior and disease transmission.

DL models like CNNs and BP ANN have proven to be effective in evaluating epidemiological data to forecast disease transmission dynamics. These models can pinpoint important risk factors, such as environmental variables, socioeconomic position, and human behaviour, that contribute to the spread of disease (Xu et al. 2013; Parsons and Banitaan 2021). These findings are critical for improving public health initiatives and developing focused interventions to lessen the burden of neglected vector-borne illnesses.

DL algorithms also help in identifying disease vectors with high accuracy. These algorithms enable quick and accurate species identification by utilizing image processing techniques, which is essential for efficient vector surveillance and control strategies (Khalighifar et al. 2019).

Integrating DL methodologies with geographic information systems (GIS) amplifies our capacity to comprehend and address health crises. Researchers can pinpoint regions at high risk of transmission and prioritize interventions by evaluating geographical data and mapping disease hotspots. Public health officials can better distribute resources with this geographically informed strategy, increasing the effectiveness of control efforts and lowering the disease burden among vulnerable groups (Hammad et al. 1996).

Furthermore, DL-powered smartphone apps like AutoML present a viable way to quickly diagnose diseases in environments with limited resources (Milad et al. 2023). These apps allow frontline healthcare workers to use basic mobile devices to diagnose diseases like trachoma with high accuracy, democratizing access to artificial intelligence in healthcare. These technologies aid in the prevention and control of neglected VBDs, especially in underprivileged populations, by enabling early detection and timely treatment.

Utilizing DL in forecasting, diagnostics, epidemiology, entomology, drug discovery, and public health offers a rare chance to transform our knowledge and management of underappreciated vector-borne illnesses. This multidisciplinary strategy makes use of cutting-edge computer techniques to evaluate large, complicated data sets, pinpoint important risk factors, and create focused solutions. Though there are several possible advantages, there are still several areas that require detailed investigation and a well-planned study. Creating DL models that can handle increasing complexities and are well suited to the subtleties of vector-borne illness transmission is one such topic. Researchers can improve these models' accuracy and prediction power by adding other data sources, such as environmental sensor data or genetic sequencing data, which will ultimately result in more efficient management measures.

Additionally, interdisciplinary cooperation is required to address the particular difficulties brought about by neglected vector-borne illnesses in various geographical locations and socioeconomic circumstances. Collaborations between researchers, medical professionals, and community organizations may be necessary to gather and evaluate data, create culturally competent interventions, and put long-term control measures into place. Further research on the moral ramifications of applying DL technology in public health contexts is also possible, especially with regards to privacy, equity, and accountability concerns. By filling in these research gaps and utilizing interdisciplinary integration to its fullest, we may significantly improve public health outcomes worldwide and lessen the burden of neglected VBDs.

Table 9 provides a comprehensive overview of the interdisciplinary approaches reviewed, highlighting their significance in disease prediction, diagnostic techniques, entomology, epidemiological studies, drug discovery and treatment, and public health interventions, showcasing their potential relevance and application in our study. Even though there is a lot of potential for integrating DL with public health, entomology, and epidemiology, there are a few challenges that

Table 9 Interdisciplinary approaches reviewed

Theme	Details
Disease prediction and forecasting	Transmission risk modelling (Xu et al. 2013); Prediction of unresponsive cases (Bamorovat et al. 2021); Predictive modelling of parasite load (Torrecilha et al. 2017)
Diagnostic techniques	Raman spectroscopy-based diagnosis (Zheng et al. 2018); Classification of ultrasound images (Wu et al. 2022); Lesion segmentation (Xin et al. 2020); Ultrasound-based diagnosis (Yang et al. 2023); Computed Tomography-based detection (Wang et al. 2023); Smartphone-based diagnosis (Milad et al. 2023)
Entomology	Vector identification of snails and Schistosoma (Tallam et al. 2021); Vector identification of snails and parasites (Liu et al. 2019); Chagas vector identification (Khalighifar et al. 2019); Chagas vector identification (Parsons and Banitaan 2021)
Epidemiological Studies	Neural network-based epidemiology modelling (Hammad et al. 1996)
Drug Discovery and Treatment	Trypanocidal drug identification (Guerra et al. 2013); Development of chitosan microparticles (Leonardi et al. 2009)
Public Health Interventions	Epidemiology modelling (Hammad et al. 1996); Smartphone-based diagnosis (Milad et al. 2023)

need to be resolved. These include addressing regional differences in illness occurrence, making DL technology accessible in environments with limited resources, and requiring bigger and more varied datasets. Getting past these obstacles will be essential for achieving the maximum benefit of interdisciplinary integration in the fight against neglected vector-borne illnesses and improving public health outcomes worldwide.

4.4 Research question 4

Novel approaches and improvements in global health outcomes can be attained in neglected VBDs like cutaneous leishmaniasis, hepatic echinococcosis, trachoma, schistosomiasis, and Chagas disease. ANNs and CNNs are two examples of DL models that have demonstrated promise in forecasting various diseases and identifying critical risk variables that contribute to the spread of disease (Xu et al. 2013; Tallam et al. 2021; Liu et al. 2019). When it comes to identifying intricate image collections associated with VBDs, such as pictures of snails and parasites, these algorithms have proven to be highly accurate (Tallam et al. 2021; Liu et al. 2019). Furthermore, DL algorithms have surpassed conventional methods in the high accuracy rates with which they have automated the detection of disease vectors as shown by Khalighifar et al. (2019) and Parsons and Banitaan (2021).

Understanding local conditions and disease epidemiology is crucial since the application of DL in forecasting VBDs can vary depending on geographical changes, demographic trends, and regional disparities. DL can help us better understand the dynamics of disease transmission and boost disease surveillance and management by integrating varied data sources and creating more sophisticated predictive modeling algorithms, as shown by Xu et al. (2013), Tallam et al. (2021) and Khalighifar et al. (2019). Furthermore, DL has demonstrated potential for improving drug formulations and resolving issues with drug solubility and absorption in the context of drug development for ailments such as Chagas disease (Leonardi et al. 2009)

Research on neglected VBDs could potentially address important issues, including quick disease detection, precise diagnosis, and efficient treatment monitoring, by incorporating DL. DL makes it possible to create prediction models that are adapted to local variations and demographic trends by utilizing a variety of data sources and cutting-edge algorithms. This improves disease surveillance and management. Additionally, DL algorithms enable the automation of diagnostic processes, which promotes prompt interventions and resource allocation-particularly in situations with limited resources where access to healthcare professionals may be restricted.

In conclusion, by enhancing disease prediction, comprehending transmission dynamics, and maximizing therapeutic development, DL in neglected VBDs research offers creative solutions and advances in global health outcomes. These developments could improve disease prevention tactics and help public health initiatives in areas of the world where VBDs are prevalent.

The study covers several neglected VBDs, including leishmaniasis, echinococcosis, trachoma, schistosomiasis, and chagas disease. However, it excludes some important illnesses due to limited available DL research studies, such as lymphatic filariasis, hydatidosis, taeniasis, cysticercosis, onchocerciasis, and mycetoma. These diseases present significant health challenges in various regions but are not extensively addressed, although they have been targeted for elimination in recent times. DL can also be beneficial in a variety of research and disease management contexts. DL algorithms can be applied to improve treatment plans, anticipate disease transmission patterns, and automate diagnosis. A more thorough grasp of the possibilities of DL in neglected VBD research can be attained by bringing these diseases into conversation.

DL applications in neglected VBDs study fields such as cutaneous leishmaniasis, hepatic echinococcosis, trachoma, schistosomiasis, and chagas disease could result in creative solutions and improvements in global health outcomes. These diseases pose serious health concerns in different parts of the world. By evaluating enormous volumes of clinical data and finding trends that could enhance patient outcomes, DL algorithms, for instance, might be utilized to enhance treatment regimens (Xu et al. 2013; Tallam et al. 2021; Liu et al. 2019).

Furthermore, by combining several data sources, including demographic trends, environmental factors, and vector populations, DL models can be used to predict disease transmission patterns, leading to more efficient disease surveillance and control (Xu et al. 2013; Tallam et al. 2021; Khalighifar et al. 2019).

Additionally, it has been shown that DL algorithms may automate disease detection through image analysis; this could be especially helpful in places where access to medical personnel is limited (Wu et al. 2022; Xin et al. 2020; Wang et al. 2023; Milad et al. 2023). Researchers can improve disease surveillance and management by utilizing DL in these neglected VBDs to create advanced predictive models that are adapted to regional differences and demographic trends. Also, DL can help create novel medicine formulations by enhancing the prediction of the bioavailability of drugs, which will result in better therapies for Chagas disease (Leonardi et al. 2009).

Finally, neglected vector-borne illnesses that can be treated using DL applications include cutaneous leishmaniasis, hepatic echinococcosis, trachoma, schistosomiasis, and Chagas disease. DL can completely transform global health outcomes in several domains by improving disease prediction, comprehending transmission patterns, and optimizing treatment techniques. A more thorough grasp of the possible advantages and challenges of using DL in neglected VBD research will require concerted efforts for interdisciplinary approaches leading to disease elimination shortly.

5 Conclusion

This comprehensive review summarizes the significant advances made possible by DL approaches in combating neglected VBDs, emphasizing their potential to transform disease monitoring, diagnosis, and management tactics. This work sheds light on the adaptability and usefulness of DL models across a range of neglected VBDs, including schistosomiasis, Chagas disease, Leishmaniasis, Echinococcosis, Trachoma, and others, through a comprehensive investigation of a large body of literature.

The findings demonstrate the effectiveness of DL algorithms in predictive modeling, species identification, diagnosis precision, and therapy optimization for neglected VBDs. Specifically, CNNs and ANNs have emerged as superior alternatives to traditional methods in disease prediction, vector identification, and parasite classification, presenting novel solutions to persistent issues in disease control and management. Also, the interdisciplinary application of DL to subjects such as epidemiology, entomology, and public health has greatly improved our understanding of disease dynamics and permitted focused interventions. DL-powered systems have intriguing opportunities for automating disease surveillance and diagnosis, especially in resource-constrained regions where access to healthcare experts is limited. The development of DL-powered smartphone applications enhances disease detection and diagnosis, thereby increasing healthcare accessibility. The review also emphasizes DL's ability to forecast disease transmission risks, identify disease vectors, classify parasites, and optimize treatment techniques for a wide range of neglected VBDs.

While this study focuses on specific diseases, it is important to recognize that DL has the potential to treat a broader variety of neglected VBDs, including lymphatic filariasis, hydatidosis, taeniasis, cysticercosis, onchocerciasis, and mycetoma. There is a limitation on the application of DL to these diseases in this review due to the unavailability of research studies. Furthermore, the potential of DL should be expanded to include vector resistance prediction, health data integration, and community engagement, which could foresee the formation of drug and parasite resistance, integrate diverse data for a more thorough understanding, and provide appealing community education tools. Expanding DL's scope to include these areas will improve interventions for VBDs and public health. Advocating for a broad research agenda and discourse is critical for realizing the full potential of DL in eliminating neglected VBDs and improving global health outcomes.

The study highlights the potential of DL for transforming disease monitoring, diagnosis, and management. By enabling targeted and efficient techniques, DL can reduce the global burden of neglected VBDs, contributing significantly to the elimination of these diseases. Continued study and collaboration are required to fully harness the power of DL and expand its influence on neglected VBDs worldwide.

Funding No funding was received for conducting this review.

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

Bamorovat M, Sharifi I, Rashedi E, Shafiian A, Sharifi F, Khosravi A, Tahmouresi A (2021) A novel diagnostic and prognostic approach for unresponsive patients with anthroponotic cutaneous leishmaniasis using artificial neural networks. PLoS ONE 16(5):e0250904. https://doi.org/10.1371/journal.pone.0250904

Ghosh S, Dasgupta R (2022) Machine learning and neglected tropical diseases. Mach Learn Biol Sci. https://doi.org/10.1007/978-981-16-8881-2_12

Guerra A, Gonzalez-Naranjo P, Campillo NE, Cerecetto H, Gonzalez M, Paez JA (2013) Artificial neural networks based on codes descriptors in pharmacology: identification of novel trypanocidal drugs against chagas disease. Curr Comput Aided Drug Des 9(1):130–140

Hammad T, Abdel-Wahab M, DeClaris N, El-Sahly A, El-Kady N, Strickland G (1996) Comparative evaluation of the use of artificial neural networks for modelling the epidemiology of schistosomiasis mansoni. Trans R Soc Trop Med Hyg 90(4):372–376. https://doi.org/10.1016/s0035-9203(96)90509-x

Kaur I, Sandhu AK, Kumar Y (2021) Analyzing and minimizing the effects of vector-borne diseases using machine and deep learning techniques: a systematic review. In: 2021 sixth international conference on image information processing (ICIIP), vol. 6, pp. 69–74. https://doi.org/10.1109/ICIIP53038.2021.9702662

Khalighifar A, Komp E, Ramsey JM, Gurgel-Gonçalves R, Peterson AT (2019) Deep learning algorithms improve automated identification of chagas disease vectors. J Med Entomol 56(5):1404–1410. https://doi.org/10.1093/jme/tjz065

Leonardi D, Salomón CJ, Lamas MC, Olivieri AC (2009) Development of novel formulations for chagas' disease: optimization of benznidazole chitosan microparticles based on artificial neural networks. Int J Pharm 367(1–2):140–147. https://doi.org/10.1016/j.ijpharm.2008.09.036

Liu ZYC, Chamberlin AJ, Shome P, Jones IJ, Riveau G, Ndione RA, Bandagny L, Jouanard N, Van Eck P, Ngo T, et al. (2019) Identification of snails and parasites of medical importance via convolutional neural network: an application for human schistosomiasis. bioRxiv: 713727. https://doi.org/10.1101/713727

Milad D, Antaki F, Robert MC, Duval R (2023) Development and deployment of a smartphone application for diagnosing trachoma: leveraging code-free deep learning and edge artificial intelligence. Saudi J Ophthalmol 5:55. https://doi.org/10.4103/sjopt.sjopt_106_22

Parsons Z, Banitaan S (2021) Automatic identification of chagas disease vectors using data mining and deep learning techniques. Eco Inform 62:101270. https://doi.org/10.1016/j.ecoinf.2021.101270

Shaikh SG, Kumar BS, Narang G, Pachpor NN (2023) Diagnosis of vector borne disease using various machine learning techniques. Int J Intell Syst Appl Eng 11(4s):517–526

Sharma DK, Chatterjee M, Kaur G, Vavilala S (2022) Deep learning applications for disease diagnosis. In: Deep learning for medical applications with unique data, pp. 31–51. https://doi.org/10.1016/b978-0-12-824145-5.00005-8

Tallam K, Liu ZYC, Chamberlin AJ, Sokolow SH, De Leo GA (2021) Identification of snails and schistosoma of medical importance via convolutional neural networks: a proof-of-concept application for human schistosomiasis. Front Public Health 9:642895. https://doi.org/10.3389/fpubh.2021.642895

Torrecilha RBP, Utsunomiya YT, da Silva Batista LF, Bosco AM, Nunes CM, Ciarlini PC, Laurenti MD (2017) Prediction of lymph node parasite load from clinical data in dogs with leishmaniasis: an application of radial basis artificial neural networks. Vet Parasitol 234:13–18. https://doi.org/10.1016/j.vetpar.2016.12.016

Wang Z, Bian H, Li J, Xu J, Fan H, Wu X, Cao Y, Guo B, Xu X, Wang H et al (2023) Detection and subtyping of hepatic echinococcosis from plain CT images with deep learning: a retrospective, multicentre study. Lancet Dig Health 5(11):e754–e762. https://doi.org/10.1016/s2589-7500(23)00136-x

Wu M, Yan C, Wang X, Liu Q, Liu Z, Song T (2022) Automatic classification of hepatic cystic echinococcosis using ultrasound images and deep learning. J Ultrasound Med 41(1):163–174. https://doi.org/10.1002/jum.15691

Xin S, Shi H, Jide A, Zhu M, Ma C, Liao H (2020) Automatic lesion segmentation and classification of hepatic echinococcosis using a multiscale-feature convolutional neural network. Med Biol Eng Comput 58:659–668. https://doi.org/10.1007/s11517-020-02126-8

Xu JF, Xu J, Li SZ, Jia TW, Huang XB, Zhang HM, Chen M, Yang GJ, Gao SJ, Wang QY et al (2013) Transmission risks of schistosomiasis japonica: extraction from back-propagation artificial neural network and logistic regression model. PLoS Negl Trop Dis 7(3):e2123. https://doi.org/10.1371/journal.pntd.0002123

Yang Y, Cairang Y, Zhou J, Zhang L, Qi B, Ma S, Tang L, Xu D, Bu L, Bu R et al (2023) Ultrasound identification of hepatic echinococcosis using a deep convolutional neural network model in china: a retrospective, large-scale, multicentre, diagnostic accuracy study. Lancet Dig Health 5(8):e503–e514

Yu Z, Wang K, Wan Z, Xie S, Lv Z (2023) Popular deep learning algorithms for disease prediction: a review. Clust Comput 26(2):1231–1251. https://doi.org/10.1007/s10586-022-03707-y

Zheng X, Lü G, Du G, Yue X, Lü X, Tang J, Mo J (2018) Raman spectroscopy for rapid and inexpensive diagnosis of echinococcosis using the adaptive iteratively reweighted penalized least squares-kennard-stone-back propagation neural network. Laser Phys Lett 15(8):085702. https://doi.org/10.1088/1612-202X/aac29f

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Knowledge and Information Systems
https://doi.org/10.1007/s10115-024-02149-9

REGULAR PAPER

Deep learning-empowered intrusion detection framework for the Internet of Medical Things environment

Prashant Giridhar Shambharkar[1] · **Nikhil Sharma**[1]

Received: 5 February 2024 / Revised: 22 April 2024 / Accepted: 20 May 2024
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2024

Abstract
The fusion of Internet of Things (IoT) technology into healthcare, known as the Internet of Medical Things (IoMT), has significantly enhanced medical treatment and operational efficiency. Real-time patient monitoring (RPM) and remote diagnostics enabled by IoMT allow doctors to treat more patients effectively and save lives. However, healthcare devices' interconnected nature makes them vulnerable to cyber-attacks, threatening patient privacy and security. Ensuring the security and accuracy of patient health data is paramount, as any tampering could have life-threatening consequences, especially in emergency situations. To address these challenges, this research focuses on developing robust security models to secure patient data in IoMT networks while meeting the growing demand for efficient healthcare services. Artificial intelligence (AI)-based technologies such as machine learning (ML) and deep learning (DL) have the potential to be employed as the methodology for intrusion detection. The goal of this research is threefold: firstly, the linear support vector machine (LinSVM) model; secondly, the convolutional support vector machine (ConvSVM) model; and finally, the categorical embedding (CatEmb) model, which have been proposed to overcome the issue of security in a network. This article offers the CatEmb model as the first effort to use a DL-based embedding approach to recognize intrusion in the IoMT environment, utilizing patient biometric and network traffic flow data. Our experimental results show the efficacy of the proposed DL models, with the LinSVM achieving a training accuracy of 99.78%, ConvSVM reaching 99.98%, and CatEmb achieving 99.84%. These models outperform existing methodologies by 2.61% in detecting network intrusions, as demonstrated through metrics such as detection rate and F1-score. Furthermore, the proposed approaches are thoroughly compared with the existing state-of-the-art studies.

Keywords Artificial intelligence (AI) · Cyber-attacks · Deep learning (DL) · Healthcare · Internet of Medical Things (IoMT) · Intrusion detection system (IDS) · Machine learning (ML) · Security

✉ Nikhil Sharma
nikhilsharma1694@gmail.com

Prashant Giridhar Shambharkar
prashant.shambharkar@dtu.ac.in

[1] Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

1 Introduction

The latest advancements and innovations in developing IoT systems have revolutionized the industry as internet-connected devices are increasing rapidly. According to Gartner, by 2025, there will be 27 billion IoT devices, which will be twice as many as the number of Internet-connected intelligent devices today [1]. It revolutionized the healthcare industry by integrating intelligence accomplishment into medical devices. Nowadays, healthcare monitoring systems have been designed using IoT as they have low-cost sensors and consume low power. Recently, these sensors have been used to expedite the patient's remote monitoring. It also lessens the requirement for doctors to be physically present every time. The fusion of IoT into medical devices is known as the Internet of Medical Things (IoMT). Nearly 30% of the IoT devices market has been covered by IoMT [2]. The modern developments and trends in wireless communication and IoT have widely supported an immense range of healthcare applications like patients' health monitoring, early detection and diagnosis of various diseases, remote healthcare services [3], and healthcare emergencies [4]. Adapting practical and secure techniques for life-threatening (critical) emergency cases can lessen the dependency and need on caregivers and cheapen medical service costs [5]. Developing intelligent decisiveness systems can facilitate early cures and possibly save people's lives. Wearable devices are employed to examine the patient health of the community continuously. So, the caregivers can impart communication regarding monitoring and diagnosis services to the people belonging to innovative societies [6]. It may create a severe problem if these systems are affected by any security threats because this gives rise to a contravention of patients' health data and medical and privacy issues, which may also result in the death of those patients in extreme cases [7].

According to the Cynerio report, most IoT-connected healthcare equipment may clasp critical susceptibility [8]. The unexpected disclosure of these gadgets in public may give rise to uncertainty risks and aid the attackers in accessing the patient's records. The adversaries may also use the known susceptibility and advanced persistent threats (APT) to deal with the victim device. This might endanger people's lives. Therefore, security must be considered a top concern in an IoMT-based healthcare monitoring system. Using different methodologies and techniques, the detection and prevention of attacks can be implemented. Other attacks, namely denial of services (DoS), malicious network traffic injection, and the man in the middle (MITM), are accomplished to compromise and attack the IoMT networks [9]. Various techniques like intrusion detection systems (IDS), vulnerability management, log monitoring, end device monitoring, prevention systems, and threat intelligence are used for detection and mitigation. IDS is a modern technology among these systems that recognizes various network attacks and security issues in IoMT.

In IDS, signature-based rules, security policies, or network traffic anomalies are executed to recognize security attacks in IoT-based networks. As the attackers used modern hacking methods and updated attacking strategies, these traditional policies were inadequate to secure the system. For example, the attacker can easily evade security policies by performing network reconnaissance and configuring firewalls and routers by reverse engineering the network devices. Researchers are investigating ML and DL approaches as they have become vital in offering solutions for developing efficacious attack detection systems. Exposure to processing and computing capabilities permits researchers to use ML and DL approaches as they give better results in precisely forecasting attack events. It is very important to secure IoMT systems as medical data are very sensitive and can be affected by cyber-attacks. The significance and challenges associated with these systems are explained as follows:

- *Patient privacy and confidentiality:* Healthcare data, such as treatment plans, patients' medical reports, documents, and diagnoses, are highly sensitive data and must be kept confidential. Security breaches may result in identity theft, privacy violation, or even blackmail.
- *Patient Safety:* IoMT devices, such as pacemakers and insulin pumps, frequently have a direct impact on patient treatment. Any unauthorized access or manipulation of these devices could endanger patient safety or even lead to loss of life.
- *Data integrity:* It is essential to maintain the integrity of medical records for proper diagnosis and treatment. Medical record tampering could lead to inaccurate treatment decisions and endangering patient health.
- *Trust in the healthcare system:* Trust plays a vital role in the healthcare system. Patients must trust that healthcare suppliers are protecting their data from different security issues. If patients no longer trust their healthcare providers, they may withhold information or even stop coming in for treatment completely.
- *Sensitive Medical Data:* IoMT devices gather an immense amount of personal health data, such as vital signs, diagnosis, and specifics of treatment. This private information might be made public through a security breach, which could result in identity theft, intolerance based on medical history, or even blackmail.

The challenges of securing IoMT systems by considering the sensitivity of medical data and cyber-attacks are explained as follows:

- *Complex environment:* It is inherently difficult to secure IoMT systems as they comprise a large number of interconnected devices, networks, and platforms. Each component may have its own vulnerabilities which requires comprehensive security measures.
- *Resource constraints:* Many healthcare companies struggle with a lack of funding, technical expertise, and cybersecurity experts. Ensuring the security of IoMT systems can be difficult because it requires advanced security technologies and training staff to address the increasing cyber threats.
- *Evolving cyber threat landscape:* The cyber threats that specifically target IoMT systems are continuously developing, as attackers are using more advanced methods to overcome security measures. Maintaining high security for IoMT (Internet of Medical Things) devices within healthcare environments presents a continuous challenge. Healthcare industries struggle to keep pace with the evolving landscape of cyber threats and vulnerabilities associated with these devices. This requires proactive security measures to mitigate potential risks and safeguard sensitive patient data.
- *Compromised Medical diagnosis:* Tampering IoMT data as a result of a cyber-attack could lead to severe consequences. Malicious individuals possess the ability to modify prescription dosages, manipulate vital signs, or interfere with device performance, thereby resulting in inaccurate diagnoses, unsuitable treatment strategies, or even life-threatening circumstances.
- *Disrupted healthcare operations:* The entire healthcare industry might be severely impacted by a successful cyber-attack on IoMT systems. Hospitals could become unable to deliver appropriate medical care if they lose access to essential patient data. Medical information and other linked devices could also be at risk if the attack spreads throughout the network.
- *Heterogeneity:* IoMT networks consist of a wide variety of devices made by different manufactures, each with a different level of security. This makes it difficult to implement a one-size-fits-all intrusion detection approach.

- *Real*-time data: Data streams are continuously produced by IoMT systems. Effective analysis of this real-time data is necessary for intrusion detection in order to identify abnormalities quickly and prevent potential harm.

1.1 Motivation

The research was mainly carried out on traditional network-based datasets containing network flow traffic samples only. In [10], an intelligent IDS framework has been developed using ML and DL techniques for identifying cyber-attacks in traditional networks. However, these approaches are inappropriate in the IoMT as, in today's scenario, different IoT-enabled health sensors are linked to the internet. Even for evaluating attack detection, the traditional network datasets are unsatisfactory in the IoMT environment. Most research emphasizes evaluating network traffic to recognize attacks in intelligent healthcare applications. In healthcare applications, patient biometric data are essential to understand their health conditions better. There is a connection between unexpected falls in the attack impacting the network and patients' sensing information, affecting healthcare records' availability, confidentiality, and integrity. Therefore, in this paper, a dataset [11] containing both patient biometrics and network traffic samples has been considered to forecast the attack incident and evaluate the two diverse data types when the attack incident takes place. The security violation in the network or IoMT devices may cause intervention in obstructed communication, loss of information related to patient's health, and disease diagnosis. As the attackers are so expert and skilled, the security risks to IoMT devices are increasing daily. To our knowledge, we contemplate this research as the first effort to use a DL-based embedding approach to recognize intrusion in the IoMT environment that has bypassed the current state-of-the-art performance.

1.2 Contribution

The significant contribution of this work is explained as follows:

- This study presents DL models, i.e., linear support vector machine (LinSVM), a collaborative model: ConvSVM (combination of convolutional neural network (CNN) and support vector machine), and categorical embedding (CatEmb) to identify network-based assaults in the Internet of Medical Things (IoMT). This research provides the CatEmb model as the first-ever method that employs a DL-based embedding approach to recognize intrusion in the IoMT environment using a WUSTL-EHMS-2020 real-time dataset.
- The model's performance evaluation is carried out using a binary classification method that distinguishes innocuous network traffic from man-in-the-middle assaults, which include spoofing and Internet of Things (IoT) data injection attacks.
- Various preprocessing and feature engineering techniques that include dropping irrelevant columns with the help of a covariance matrix have been demonstrated, as these features may not significantly affect the detecting process. This helps build an effective model that takes less memory utilization and inference time in detecting attacks in the network. The input dataset is balanced using a data augmentation approach to provide enough data availability for both the regular and attack samples. Segregate the dataset features into categorical and continuous columns and implement well-known feature engineering approaches like standard scaler, a power transformer for continuous columns, and ordinal encoders for categorical columns.

- The proposed models have been contrasted with the existing techniques explored on a similar real-time and benchmark dataset, showing that our models are superior with the highest accuracy, precision, recall, and F1-score for detecting attacks in minimum inference time.

The rest of the paper is systematically organized as follows: Sect. 2 highlights the literature review, which covers the capability of ML and DL models on different network flow datasets; Sect. 3 explains the architecture and workflow of the proposed methodology for intrusion detection in the IoMT ecosystem, along with in-depth discussion of data preprocessing, hyperparameter tuning of ML models, and proposed DL models; Sect. 4 illustrates about the experimental setup, testbed and dataset description, experimental work, and comprehensive performance evaluation of proposed models with the existing work followed by the conclusion in Sect. 5.

2 Literature review

This section provides a comprehensive study of different IDS frameworks proposed by researchers to identify the attacks in the IoT environment using ML and DL approaches and highlights the need for security. Intruders may harm the IoT network in the absence of IDS. Therefore, IDS infrastructure plays a vital role in eliminating such research gaps. Here, multiple frameworks have been explored, which are discussed as follows:

Clifton et al. [12] explained the ML technique's role in health monitoring systems. These techniques have been used for controlling and managing false alerts while revealing severe health issues. The data used in their experiment combine the patient's clinical observations to provide quick alerts in an envisaged emergency. This work has been conducted at Oxford University Hospital. Fotouhi et al. [13] developed a healthcare monitoring framework comprising three integrants, i.e., a gateway, an access point (AP), and a coordinator. The coordinator is the node attached to the patient's body for gathering information related to the patient's health using sensors. Static nodes, called access points, are located in the room's walls and by sensors that use communication protocols (like 6LoWPAN, BLE, or ZigBee). The information gathered by the APs gets forwarded to the gateway. Then, this information gets transferred to the cloud server using the internet. Without proper testing and explanation, some techniques have been proposed in this system for securing the data. Baburaj and Rani [14] introduced a cloud-based healthcare platform that uses an SVM (support vector machine) approach to forecast patients' situations and envisioned diseases. No unauthorized users are permitted in the system.

A blockchain-based healthcare system framework has been designed by Chakraborty et al. [15], which helps overcome the problems associated with the traditional healthcare system related to the security issues of the records created during patient treatment. The framework has been set up to supervise the treatment process from beginning to end. Alabdulatif et al. [16] developed a cloud-based smart prediction framework. This system was based on the fully homomorphic encryption (FHE) approach and comprises three blocks, i.e., the intelligent community resident, cloud storage, and smart prediction model. The data is gathered and dispatched to the cloud storage repository system in the first block. The encrypted data gets amassed in the second block. The third block has been used to detect abnormal changes like attacks without decrypting data. A secured anticipating approach based on Holt's linear trend method has been developed to predict bizarre changes in patient's vital signs, which helps

detect different chronic diseases. The author also introduced a novel parallel technique of Holt's method for improving the effectiveness of the FHE model.

The development of network intrusion detection systems (NIDS) has resulted in an increasing adoption of artificial intelligence (AI) techniques, particularly ML and DL, in recent years. Extensive research has been conducted on both ML and DL techniques, with a significant trend toward DL. The reason for this choice might be traced to the progress made in graphical processing units (GPUs), which have facilitated faster computations required for deep learning methods. As a result, scholars are progressively using deep learning algorithms to suggest more accurate security measures for Internet-of-Things (IoT) networks, which handle substantial amounts of unprocessed data. Deep learning algorithms possess a high level of proficiency in analyzing complex patterns due to their deep design, which enables them to effectively distinguish between regular and atypical network data.

2.1 ML-based IDS frameworks for attack detection

This subsection discussed different ML-based frameworks for solving security-related problems in IoT networks. A security system has been suggested by Zhang et al. [17] that applied the RF (random forest) technique to detect anomaly traffic on the KDD-1999 dataset. With a 1% false positive rate (FPR), this technique achieves 95% accuracy as an anomaly detector. This dataset is employed to test anomaly detection algorithms. It is a generic Knowledge discovery and data mining dataset. Since 1999, this dataset has been used in many competitions. Rao and Swathi [18] employed the Indexed Partial Distance Search k-nearest Neighbor (IPDS-KNN) technique to assess various attacks. It achieved 99.6% accuracy performance. Shapoorifard et al. [19] use the k-nearest neighbor (KNN) approach and mainly emphasize decreasing the false alarm rate (FAR) and achieving 85.2% accuracy. Tao et al. [20] introduced a SecureData scheme that delivers privacy and security to patients' private records. The FPGA (field programmable gate array) platform is used to optimize the KATAN secret algorithm, which is implemented for secure communication. To overcome the attack detection problem in IoMT, Yaacoub et al. [21] discussed different types of ML-based privacy and security solutions. But, according to the author, there is still a need to introduce an effective IDS system to detect attacks. To analyze attacks in the smart hospital, an ensemble classifier IDS has been developed by Saba [22]. The decision tree (DT) technique attained 93.2% accuracy performance in categorizing the cyber-attacks in the KDDcup-99 dataset. This dataset was created in the traditional network without adding IoT device traffic. A mobile agent-based IDS has been introduced by Thamilarasu et al. [23] to identify both network and device-based attacks in the IoMT environment. The simulation-generated datasets were tested using ML and regression techniques. Using the DT technique in the evaluation process, the device and network-level intrusion detection achieved an accuracy of 97.93% and 99.8%, respectively. Kumar et al. [24] experimented in three stages. The author introduced an RF, naïve Bayes (NB), and DT ensemble in the first stage. In the second stage, XGBoost was applied to categorize regular and attack network records. In the third stage, to classify the attacks in the IoMT environment, the developed model was applied to the ToN-IoT dataset, which attained 96.35% accuracy. The industrial IoT network setup has been used to create this dataset using Modbus weather sensors. In the IoMT environment, these sensors are not commonly employed. Therefore, the above data could not be appropriate for identifying network attacks.

Radoglou et al. [25] developed an Intrusion detection and prevention system (IDPS) to identify and prevent various cyber-attacks against communication protocols like Modbus/TCP and HTTP, broadly used by e-healthcare services. Here, EHR uses HTTP, whereas IoMT uses Modbus/TCP protocol. The proposed IDPS can retrain ML techniques and test itself using an active learning approach. The CIC-IDS2017 dataset was employed in this experiment to analyze the functioning of ML techniques on the HTTP network dataset. DT classifier achieved an accuracy of 96.44% in categorizing network attacks. In comparison, RF attained 94.45% of the accuracy on the Modbus dataset. Zachos et al. [26] introduced a systematic and potent anomaly-based IDS (AIDS) for the IoMT environment. The three features, i.e., gateways, IoT devices, and network traffic features, were combined to devise a unique feature set. Various ML techniques have been applied to identify deviations in the gathered malicious and data events in the network to enhance the functionality of attack detection. Memory consumption level attributes and CPU were considered for evaluation in IoT devices. The TON_IoT Telemetry dataset has been used in this experiment. According to the result reported by the author, KNN, RF, and DT are the most appropriate ML techniques that are employed for the central detection integrant of the introduced system.

Binbusayyis et al. [27] inspected and showed a detailed comparison of different methods like KNN, SVM, ANN (Artificial Neural Network), NB, and DT. The experiment used the Bot-IoT dataset to compare the working performance of ML methods. This dataset comprises various attack categories like denial of service (DoS), theft attacks, and distributed denial of service (DDoS) attacks. Spoofing and MITM attacks are IoMT attacks not covered in this dataset. On the tested dataset, DT attained an accuracy of 100%, and other ML techniques like SVM, NB, and KNN achieved an accuracy of 99%. Saif et al. [28] proposed a hybrid methodology known as DHOA-ANFIS, with the objective of identifying routing vulnerabilities in wireless body area networks (WBANs). The novelty of their approach involves the integration of the adaptive neuro-fuzzy inference system (ANFIS) and the distributed Hopfield neural network (DHOA) in order to improve the precision of routing attack detection. The primary aim of this research is to improve the security and dependability of wireless body area networks (WBANs) by effectively tackling unique obstacles.

As per the study, ML techniques are used to identify attacks in IoMT. However, most datasets were created without considering the IoMT environment and attacks. The results presented by the authors in their research were outstanding, as in many contributions, the ML techniques achieved an accuracy of 95%. The IoMT study considered many input features like IoT device memory, network traffic, CPU, and metric features. However, researchers did not use or mention features like patient biometric data to identify cyber-attacks in the IoMT. Many researchers explored DL techniques to classify or identify the attack in the IoMT ecosystem.

2.2 DL-based IDS frameworks for attack detection

This subsection presents the DL-based IDS framework for solving security issues in IoT networks. To forecast various attack simulations in deep brain stimulators (DBSs), Rathore et al. [29] developed a DL algorithm that efficiently identifies the pattern of attacks and alerts a patient regarding that. To identify network attacks in the IoT environment, Manimurugan et al. [30] introduced a DL-based deep belief network (DBN) algorithm that achieved an accuracy of 96%. The experiment was conducted using the proposed model on the CICIDS dataset. This dataset generation did not concentrate on IoMT network attacks. For feature

selection, Saheed and Arowol [31] used particle swarm optimization (PSO) and ML/DL-based techniques to identify cyber-attacks in the network. Researchers used the NSL-KDD dataset to analyze the functionality of the suggested approach. The introduced model attained 99.76% of the accuracy performance. This dataset was not created with the IoT environment in mind and should not be used to assess attack identification in IoMT. Awotunde et al. [32] developed a swarm neural network (SNN) method that detects intruders while transmitting the data and permits accurate and efficient assessment of medical data at the network edge. The author used the NF-ToN-IoT dataset for the experiment, which amalgamates network data, operating systems, and telemetry. The author used a deep autoencoder (DAE) to decrease the dimensions of features. The author used a deep feed-forward neural network (DFFNN) in the IoT environment to recognize the network attacks. The DAE-DFFNN model achieved an accuracy of 89%, which is superior to ML techniques such as DT and SVM, as claimed by the researcher. For identifying malware in the IoMT ecosystem, Khan and Akhunzada [33] introduced SDN (software-defined network) enabled CNN (convolutional neural network) and LSTM (long short-term memory) hybrid DL model, which attained an accuracy of 99%. However, this framework was not used as IDS to determine network attacks in the IoMT ecosystem. Nandy et al. [34] developed intelligent agent-based SNNs to detect IoMT intruders. The experiment was conducted using the proposed approach on the ToN-IoT dataset, which attained 99.5% accuracy.

The above study of DL approaches indicates that these techniques were not highly introduced to identify IoMT network attacks. Most authors only explored the network traffic dataset in their experimental work to identify the attacks in the network, as discussed in Table 1. None of the above-mentioned results contemplate the combined features of patients' biometrics with the network flow data.

3 Proposed architecture and methodology

The performance of attack identification utilizing patient biometrics and network flow dataset has been improved in this paper by introducing DL-based models to find attacks in the IoMT environment. Before explaining our proposed procedure, we first outline the architecture of IoMT IDS, as shown in Fig. 1. This architecture maintains security performance and predicts attacks using ML and DL methods. The IoMT IDS architecture combines IoT gateway, security operators, network traffic collector, IDS, patient sensor devices, and ML or DL data processing techniques, which are used to monitor different types of attacks. The patient's device sensor comprises ECG devices, a pulse rate detector, a respiration rate tracking machine, temperature sensors, and a heart rate tracking machine. The information gathered using these sensor devices gets transferred to the remote servers using IoT network protocols like Message Queuing Telemetry Transport (MQTT) and Advanced Message Queuing Protocol (AMQP). Using wired or wireless communication, the IoT gateways gathered the sensor's information and transferred that information to remote areas. The proposed techniques comprise analytical accomplishments and data processing to identify intrusion in the IoMT ecosystem. Continuous analysis and monitoring are needed at the IDS level to reduce false positives and alert the user regarding the attacker. The physical premises monitoring and remote monitoring of patients are the primary applications of the IoMT environment, which helps to cure and save patients' health in the hospitals.

Figure 2 depicts the operational structure of the IDS framework to identify attacks using ML and DL techniques in the IoMT ecosystem. Rather than analyzing network flow traffic

Table 1 ML and DL methods to identify an attack in the network

Year	Authors	Methodology	Dataset	Description	Result	Limitations
2018	Su et al. [35]	Lightweight CNN	IoTPOT	The proposed methodology was employed to recognize DDoS cyber-attacks in IoT networks	$A = 94\%$	This dataset is a combination of IoT threats. This dataset is not used to identify attacks in the IoMT environment
2018	Nguyen et al. [36]	CNN classifier	IoT Botnet	The author introduced the model that combines the CNN classifier and PSI graph for Linux IoT botnet identifications in this paper	$A = 92\%$, $P = 91\%$, $R = 97.6\%$, F1 = 94%	This dataset contained flow-based features and was only used for malware identification in IoT networks
2019	Rathore et al. [29]	Recurrent neural network (RNN)	Dataset obtained from Physionet	To forecast various attack simulations in Deep brain stimulators (DBSs), the author introduced a DL model that efficiently identifies the pattern of attacks and alerts a patient regarding that	Different low loss value, RMSE = 0.095	There was no discussion of accuracy in the paper, and the simulated attacks were not actual

Table 1 (continued)

Year	Authors	Methodology	Dataset	Description	Result	Limitations
2020	Hussain et al. [37]	LR (logistic regression), KNN, NB, RF	CICIDS2017, IoT-23, CTU-13	The authors presented the idea of a universal feature set. Different ML models were considered for classifying the attacks	$A = 100\%$, $P = 100\%$, $R = 100\%$, F1 = 100%	Low accuracy; No hyperparameter tuning was carried out during the experimental process; Techniques took high prediction time; For the Universal Classification Process, the dataset needed to be combined
2020	Farhan et al. [38]	DNN (deep neural network)	CSE-CIC-IDS 2018	To classify the attack in the network, the author proposed the DNN model	$A = 90.25\%$, $P = 65\%$, $R = 59\%$	The proposed method achieved low precision and recall values, i.e., 0.65 and 0.59, respectively. No other techniques were considered for classifying the attack. The dataset is not viewed for identifying attacks in the IoMT network

Table 1 (continued)

Year	Authors	Methodology	Dataset	Description	Result	Limitations
2020	Sarhan et al. [39]	Extra tree classifier	UNSW-BN 15, ToN-IoT, CSE-CIC-IDS2018, BoT-IoT,	In this paper, four datasets were considered to show NetFlow features. The n Probe Tool generalized the datasets into universal feature datasets. The proposed approach achieved a weighted average of 70.81%, a prediction time of 14.67 µs, and an F1-score of 0.79	$A = 90.28\%$, F1 = 94%	Prediction time is high; For classification, no other techniques have been used; High FAR rate
2021	Kumar et al. [24]	Ensemble classifiers like RF, NB, and DT	ToN-IoT dataset	The suggested models were applied to categorize the attacks in the IoMT environment	$A = 96.35\%$, DR = 99.98%	Modbus weather sensors are employed to create a dataset generally not used for the IoMT environment. It contains only network traffic data. Therefore, the dataset could not be appropriate for identifying network attacks in IoMT; The false acceptance rate (FAR) is high

Table 1 (continued)

Year	Authors	Methodology	Dataset	Description	Result	Limitations
2021	Saheed and Arowol [31]	PSO-RF	NSL-KDD dataset	The author used PSO and ML/DL-based techniques to identify malicious attacks in the network	$A = 99.76\%$, F1 = 96.45, $P = 99.75$, MCC = 99.51%	The dataset is only relevant to network traffic data and does not apply to IoMT
2021	Awotunde et al. [32]	DAE-DFFNN	NF-ToN-IoT dataset	The author presented an SNN-based method that detects intruders while transmitting the data and permits accurate and efficient assessment of medical data at the network edge	$A = 89\%$	The proposed model achieves low accuracy
2021	Khan and Akhunzada [33]	SDN-enabled LSTM and CNN hybrid DL model	–	The author proposed an SDN-enabled hybrid DL model to detect malicious attacks in the network	$A = 99\%$	This work only explored the identification of Malware attacks
2021	Nandy et al. [34]	Intelligent agent-based SNN	ToN-IoT dataset	The author proposed an intelligent agent-based SNN for detecting intruders in IoMT	$A = 99.5\%$, ER = 0.5%	The dataset is only relevant to network traffic data and does not apply to IoMT
2022	Binbusayyis et al. [27]	KNN, SVM, ANN, NB, and DT	Bot-IoT dataset	This paper showed a detailed comparison of different techniques like KNN, SVM, ANN, NB, and DT for identifying attacks in the network	$A = 100\%$, DR = 100%, FAR = 0	The dataset is relevant for network traffic data only

Table 1 (continued)

Year	Authors	Methodology	Dataset	Description	Result	Limitations
2023	Saif et al. [28]	DHOA-ANFIS	User-generated Dataset	The primary aim of this research is to improve the security and dependability of WBANs by effectively tackling unique obstacles	$A = 95.88\%$, $P = 94.44\%$, $R = 95.30\%$, FAR = 0.96	High FAR rate
2023	Saif et al. [40]	RF, MLP, CNN, SVM, DT, NB, LR	N-BaIoT	This paper showed a detailed comparison of different techniques like RF, MLP, CNN, SVM, DT, NB, and LR for identifying attacks in the network	$A = 99\%$, $P = 100\%$, $R = 100\%$, F1 = 100%	Prediction time is high
2023	Saif et al. [41]	GA-DT	NSL-kDD	This paper showed a detailed comparison of different techniques like GA-kNN, DE-DT, DE-kNN, PSO-DT, GA-DT, and PSO-kNN for identifying four different attacks in the network	$A = 99.88\%$	The dataset is only relevant to network traffic data and does not apply to IoMT

A Accuracy, *P* Precision, *R* Recall, *F1* F1-score, *MCC* Matthews correlation coefficient, *DR* Detection rate, *FAR* False alarm rate, *TPR* True positive rate, *FPR* False positive rate, *ER* Error rate

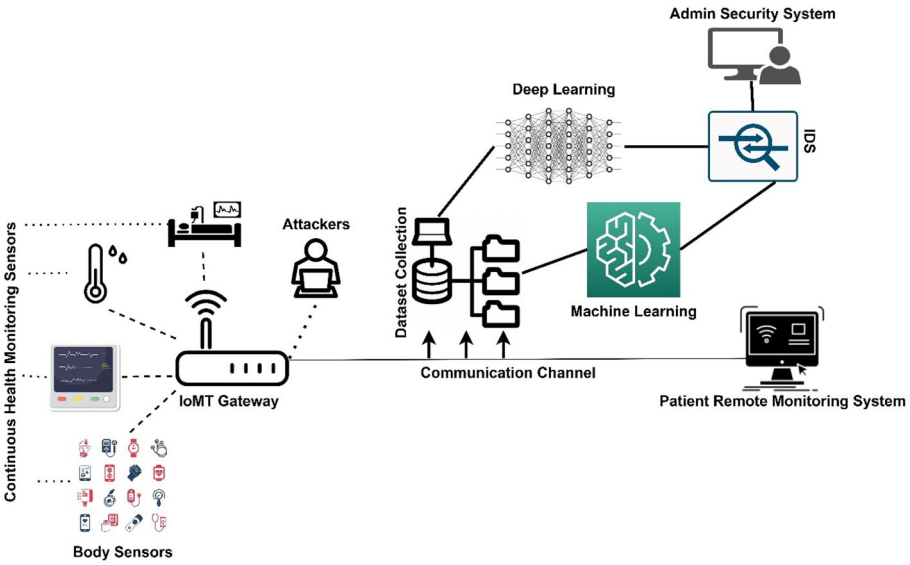


Fig. 1 ML and DL-based IDS Framework for IoMT environment

and identifying intrusion in the system, we used IoT sensing data that belongs to the patient's specific IoMT ecosystem. This IoT-based sensing data helps to detect patient biometrics anomalies and enhances the performance of attack identification whenever the intruder tries to attack the IoMT environment. With the help of both events' timestamps, i.e., patient biometric data events and network traffic events, the patient's biometric and network traffic data get combined. The final dataset presents a minority of features belonging to patient biometric data and the majority of the network traffic features.

In order to improve network-based attack detection in the Internet of Medical Things (IoMT), the framework as shown in Fig. 2 includes various DL models such as linear support vector machine (LinSVM), ConvSVM (a combination of convolutional neural network (CNN) and support vector machine), and categorical embedding (CatEmb). CatEmb is a novel embedding approach that employs deep learning techniques and leverages the WUSTL-EHMS-2020 real-time dataset for the purpose of intrusion detection. Moreover, the application of binary classification algorithms in a rigorous evaluation process allows for the differentiation between benign network traffic and potential attacks such as spoofing and IoT data injection attacks. This ensures the practical effectiveness of IoMT networks. The utilization of memory and inference time is optimized through the implementation of significant preprocessing and feature engineering approaches, such as systematic column reduction using covariance matrix analysis. Furthermore, data augmentation mechanisms are employed to achieve a balanced dataset, while feature segregation enables the utilization of conventional techniques such as standard scaler and power transformers for continuous columns and ordinal encoders for categorical columns. Finally, the optimal performance is achieved by seamlessly integrating with other components.

The computational demands of the proposed method are certainly significant, requiring substantial computational resources. To address this, we have chosen to implement the method on the centralized server of the IoMT system. By offloading the computational burden to a centralized server, we mitigate the impact of resource constraints within individual IoMT

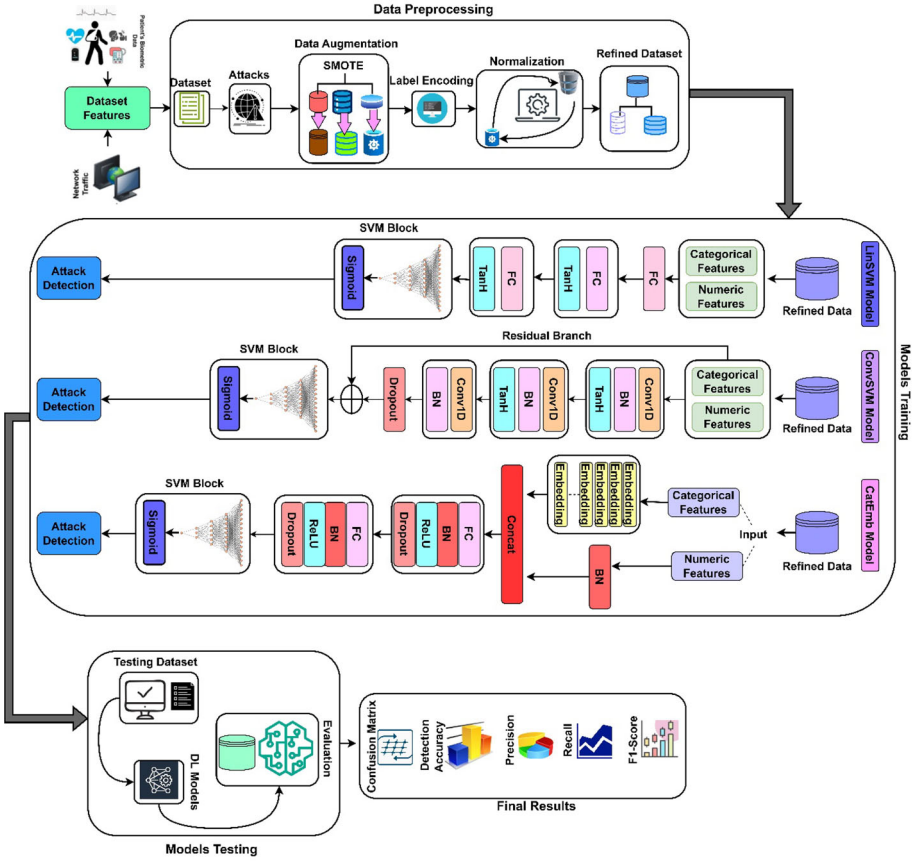


Fig. 2 Proposed work flowchart to detect IoMT attacks

devices. This design choice allows us to leverage the computational power and resources available on the server, which are typically more robust compared to those available on resource-constrained IoMT devices. Therefore, while acknowledging the high computational requirements of the method, our approach ensures that the resource constraints inherent in IoMT devices do not directly impede the implementation and functionality of the proposed method.

3.1 Data preprocessing

Data processing is a preliminary step before training and testing the ML or DL so that models can provide better results. This paper used popular techniques like standard scaler and simple imputer to preprocess the data. To begin with, standard scaler is a well-known method used for data standardization and primarily a preprocessing procedure performed before many ML or DL models to normalize the functional range of the input datasets. It is employed to scale the value distribution in such a matter that the observed value's mean is 0, and its standard

deviation is 1. Equation (1) shows the feature scaling with standard scaler [42].

$$X_{\text{FeatureScaled}} = \frac{X_{\text{FeatureCurrentValue}} - \text{mean}(X_{\text{Feature}})}{\text{StandardDeviation}(X_{\text{Feature}})} \tag{1}$$

where $X_{\text{FeatureScaled}}$ shows the scaled feature value after the successful execution of the scaling process; without any alterations, $X_{\text{FeatureCurrentValue}}$ displays the current feature value; X_{Feature} represents all of the dataset's feature columns; and the mean for each feature depicted by mean ().

In addition to the process, a dataset's missing values can be imputed using the Simple Imputer as expressed in Eq. (2). These values can be substituted with statistics like the mean, median, or most frequent value for the missing numbers, or it can use a provided constant. Here, y_{pq} represents the initial value; y'_{pq} denotes the imputed value for the elements located in the pth row and qth column; and strategy shows the imputation approaches like mean, median, most frequent value or a user-specified constant value.

$$y'_{pq} = \begin{cases} \text{strategy} & \text{if } y_{pq} = NaN \\ y_{pq} & \text{otherwise} \end{cases} \tag{2}$$

Furthermore, a power transformer [43] and a standard scaler are used to scale the data in the numeric columns. They are a family of monotonic and parametric transformations that convert skewed features into normal distributions using logarithmic transform to create data that is more Gaussian-like. It handles the modeling problem related to non-constant variance or any other type of condition where normality is required. Power Transformation is expressed by Eq. (3).

$$y(\lambda) = \begin{cases} \frac{((y+1)^{\lambda}-1)}{\lambda}, & \text{if } y \geq 0 \text{ and } \lambda \neq 0 \\ \log(y+1), & \text{if } y \geq 0 \text{ and } \lambda = 0 \\ -\frac{((-y+1)^{2-\lambda}-1)}{(2-\lambda)}, & \text{if } y < 0 \text{ and } \lambda \neq 2 \\ -\log(-y+1), & \text{if } y < 0 \text{ and } \lambda = 2 \end{cases} \tag{3}$$

Here, a parameter called λ is estimated in the fitting procedure, y indicated input data, and to make the data more Gaussian-like, the optimal value of λ is selected.

3.2 Feature engineering

It is an ML approach that uses data to develop new variables that aren't considered in the training set. It helps to modify datasets, including mutation, combination, addition, and deletion operations, to upgrade the training of the ML models and attain ameliorate accuracy and performance. Here, the feature selection approach has been applied to minimize the input variable by employing the pertinent features and eliminating the noise from the data [44]. In this paper, a co-variation matrix heatmap, which represents the linear correlations between the various variables in the dataset, has been used. It facilitates the identification of patterns and correlations between variables. The covariance between two variables, X and Y, can be determined using Eq. (4). Some columns were correlated throughout the dataset, so there is no need to consider those columns. Therefore, four columns, namely Dir, SrcAddr, DstAddr, and DtMac, were dropped by taking advice from the domain experts. After that, an ordinal encoder [45] encodes the categorical columns, i.e., Flgs and SrcMac. Each label is converted into an integer value through ordinal encoding, and the encoded data shows the order of

labels. Equation (5) is used to encode categorical column data using the ordinal encoder.

$$\mathrm{Cov}(X,\ Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \tag{4}$$

where the covariance between variables X and Y is denoted as Cov(X, Y); n represents the number of data points, X_i, Y_i; $\overline{X}$, and $\overline{Y}$ shows an average of X and Y.

$$x' = \mathrm{ord}(x) \tag{5}$$

where x' shows encoded value; ord indicates where category x is located in the given categories' parameter; x represents the original categorical value.

The dataset was portioned using distributions of 80% for the training and 20% for the testing datasets to accurately compute the ML and DL models' working performance. Further, the performance has also been evaluated for 70–30% dataset distribution and compared with 80:20% distribution output. The K-Fold cross-validation with ten folds has been used on the training dataset sample to demonstrate the range of working performance between the folds. Identifying the attacks using this model may be challenging as the dataset needed to be more balanced, with standard samples making up roughly 88% of the data. Hence, to balance the dataset in the training phase, an oversampling method, i.e., SMOTE (Synthetic Minority Oversampling Technique) [46], has been applied. The role of this technique is to balance the dataset by producing new synthetic samples for the minority class. In the SMOTE class, the fit_resample method is used from the imblearn library [47] that fits the SMOTE algorithm to the input data X and y, and returns the resampled data as expressed in Eq. (6). Here, X is a 2D array-like object representing the features of the input data, and y is a 1D array-like object, representing the target or label of the input data.

$$X_\mathrm{resampled},\ y_\mathrm{resampled} = \mathrm{smote.fit_resample}(X,\ y) \tag{6}$$

The amount of resampled data depends on the parameters passed to the SMOTE object. By default, it creates synthetic samples for the minority state class until the class distribution is balanced. Therefore, there will be equivalent data samples from the minority and majority classes. However, we can specify the amount of oversampling by setting the sampling_strategy parameter. For example, if sampling_strategy is set with a value of 0.5, then SMOTE will oversample the minority class by a factor of 0.5. Hence, there will be 50% more samples in the minority class than in the majority class. Table 2 represents the dataset before and after resampling.

Table 2 Dataset before and after resampling

IoMT dataset	Dataset before SMOTE Resampling		Dataset after SMOTE Resampling	
	Normal sample value	Attack sample value	Normal sample value	Attack sample value
WUSTL-EHMS-2020	14,272	2046	14,272	14,277

3.3 ML models & hyperparameters tuning

In this work, hyperparameter tuning to all the pre-trained ML models has been performed to improve the attack detection in the IoMT environment, which is explained as follows:

3.3.1 SVM

SVM is a linear classification and regression model that works well for real-world problems and can supervise related linear and nonlinear issues. This technique made a line or hyperplane that distinguished the data into classes [48]. To ensure that $y_i(w.x + b) \geq 1$ is always met, and the objective function must be reduced. Equation (7) represents the hyperplane function where w represents Weights, and b refers to bias.

$$H(x) = \begin{cases} +1, & \text{if } w.x + b \geq 1 \\ -1, & \text{if } w.x + b \leq 1 \end{cases} \tag{7}$$

3.3.2 DT

DT is a supervised ML technique that constantly divides the data based on a given parameter. The two entities may be used to explain the DT, viz. decision leaves and nodes. It is employed for addressing regression and classification issues [49]. The main objective of this algorithm is to develop a model that predicts the intent variable value by learning simple decision rules derived from the data attributes. In our analysis, we set the maximum depth value to 8.

3.3.3 LR

LR is a supervised ML approach for addressing classification and prediction issues. It is simple to elucidate and implement and well organized to train the data as it takes less time to classify unknown records [50]. It is used to forecast or compute the probability that the event occurs. It should be a binary target variable and provide probabilistic values ranging between 0 and 1. Equation (8) represents the LR model. In our analysis, we set the value of the random state and maximum iterations to 0 and 100, respectively.

$$q = \frac{\mathrm{e}(\mathrm{b0} + \ \mathrm{b1} * \mathrm{p})}{(1 + \ \mathrm{e}(\mathrm{b0} + \ \mathrm{b1} * \mathrm{p})} \tag{8}$$

where p is input values merged linearly using coefficient or weight values, *b0* is the intercept or bias value, *b1* represents a single input term *(p)*, and q is the predicted output.

3.3.4 KNN

It is a supervised ML approach for addressing classification and regression issues. It aims to forecast the proper class samples by measuring the distance amid all test data and training points. After that, it picks the K-Points, i.e., nearest to the test data. This technique computes the probability of the test data sample belonging to K training data classes and selects the class with the excessive probability [51, 52]. In our analysis, we choose two neighbors, set the leaf size to 100, and assign the average values from nearby neighbors as the final projected values. The closest neighbors are computed using the Euclidean distance (E), as

shown below in Eq. (9), where the number of neighbors denoted by k and the data points in the i^{th} dimension are a_i and b_i.

$$E = \sqrt{\sum_{i=1}^{k} (a_i - b_i)^2} \tag{9}$$

3.3.5 Gradient boosting classifier (Grad_Boost)

It is a collection of ML models that fuse multiple weak learning techniques, usually DT, to build a dedicated predictive model. It may be applied to address both regression and classification issues. Regularization techniques are employed to minimize the consequences of overfitting and to prevent deterioration by confirming that the fitting operation is limited [53]. It is a good technique that deals with unbalanced datasets. In our analysis, we set the value of the subsample, estimators, maximum depth, and maximum leaf nodes to 0.5, 1000, 8, and 1000 respectively. The result of H(x) for input x is defined in Eq. (10).

$$H(x) = \sum_{i=1}^{n} \alpha_i h_i(x) \tag{10}$$

where n shows number of leaves, and the value predicted for the region h_i is represented by a_i.

3.3.6 RF

A supervised ML approach is applied to address regression and classification analysis. It is based on the concept of ensemble learning, which integrates many classifiers to address challenging issues, upgrade model performance, and eliminate the problem of overfitting. With high dimensionality, this technique can manage big datasets [54]. Rather than depending on the DT, RF uses prediction from all the trees and forecasts the final result based on which predictions received the votes. In our analysis, we set the value of estimators, random state, and maximum leaf nodes to 1000, 1, and 1000, respectively.

3.3.7 XGBoost

It is a well-known supervised ML approach that is enforced to address classification and regression issues. It tries to forecast a target variable precisely by fusing multiple weak learning techniques. It is fundamentally similar to the Grad_Boost classifier, but the residual trees are constructed differently in XGBoost. The variables utilized as the roots and nodes of the residual trees are chosen by computing similarity scores between the previous nodes and the leaves [50]. In our analysis, we set the value of the estimators and maximum depth to 1000 and 8, respectively. Equation (11) represents the XGBoost model.

$$\text{ObjectiveFunction} = \text{Loss}(y_\text{true},\ y_\text{pred}) + \lambda \times \text{Complexity}(T) \tag{11}$$

where Loss (y_true, y_pred) represents the loss function that calculates the difference between the actual target values (y_true) and the expected target values (y_pred); A regularization term called complexity (T) penalizes the model's complexity; λ- referred to as the regularization parameter, determines how strong the regularization is.

Table 3 Hyperparameter tuning of ML models

ML models	Folds	Parameters optimized	Values
SVM	K-fold cross-validation = 10	Gamma	Auto
DT		Max_depth	8
LR		Random_state	0
		Max_iteration	100
KNN		N_neighbors	2
		Leaf_size	100
Grad_Boost		Subsample	0.5
		N_estimators	1000
		Max_depth	8
		Max_leaf_node	1000
RF		N_estimators	1000
		Random_state	1
		Max_leaf_node	1000
XGB		N_estimators	1000
		Max_depth	8

Table 3 depicts the hyperparameter tuning of the ML model. For the training process, all models are set with different parameter values and evaluated using 10K-fold cross-validation to lessen the variance in the findings over data partitioning and circumvent the model's overfitting.

3.4 Experimental configuration of DL models for IDS

This section discussed novel DL-based proposed models for recognizing attacks in the IoMT network. These models are trained using the different parameters, as shown in Table 4.

3.4.1 LinSVM (linear support vector machine)

In this model, we initialize the weight using Xavier Uniform [55] to ensure steady training. This helps prevent layer activation outputs from vanishing or exploding during the forward

Table 4 DL models parameters

Parameters	Values
Learning rate	0.0001
Embedding dimension	16
Epochs	200
Batch size	64
Optimizer	Adam
Activation function	ReLU, TanH

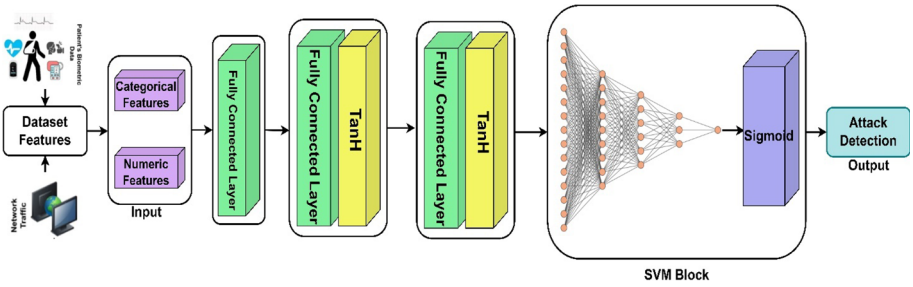


Fig. 3 Working flow of LinSVM for detecting IoMT attack

process through a DNN, and the bias terms are set to 0.01. Both categorical and numerical features were passed to block 1, containing fully connected layers (FC), in which the input gets augmented by a weight matrix in an FC, and a bias vector gets added to it. The output of block 1 is fed to block 2, containing the FC and Tanh (hyperbolic tangent) activation functions. A feed-forward neural network with several hidden layers is implemented in this class. Each hidden layer consists of a linear transformation followed by a TanH activation function. The classification between two classes is accomplished using the Tanh activation function. It has a range of between − 1 and 1. It makes learning more accessible for the next layer by centering the data and generating a mean near 0. The output of block 2 transferred to block 3, containing the same components. Then, the output from Block 3 is fed to the SVM block containing the FC layer and sigmoid activation function for the binary classification. Therefore, these modifications and nonlinear activation functions are applied in the forward technique to compute the final output. Figure 3 depicts the working flow of LinearSVM for detecting IoMT attacks.

3.4.2 ConvSVM (convolutional support vector machine)

ConvSVM is a combination of a convolutional layer and SVM. This model passed categorical and numerical features to block 1, containing a 1D convolutional layer followed by batch normalization (BN) and a Tanh activation function. Using this, hierarchical features are extracted from the input by the convolutional layers (conv1, conv2, conv3) and their related BN layers (BN1, BN2, BN3), which helps to enhance model generalization. This block applies sliding convolutional filters to 1-D input via a 1-D convolutional layer. Every value in the input vector gets multiplied by the kernel value in this operation. Using BN with a 1-D convolutional layer alleviates the model functionality by expediting the training process, employing more excellent learning rates, and eliminating internal covariate shifts. It also consumes less memory. Tanh activation function makes learning more accessible for the next layer by centering the data and generating a mean near 0. The output from Block 1 fed into Block 2, containing the same components as the earlier block. After that, the output of Block 2 was sent to Block 3, comprising a 1D convolutional layer and BN. Then, we used the dropout layer to drop noise or data value that is purposefully removed from a neural network to accelerate the processing of the model and prohibit it from overfitting. To save from vanishing gradient, we added a residual branch and concatenated it with an output of block 3. Finally, we pool the features using an adaptive average pooling layer, which was then passed to an SVM block containing an FC and the sigmoid activation function for the binary classification. Figure 4 depicts the working flow of ConvSVM for detecting IoMT attacks.

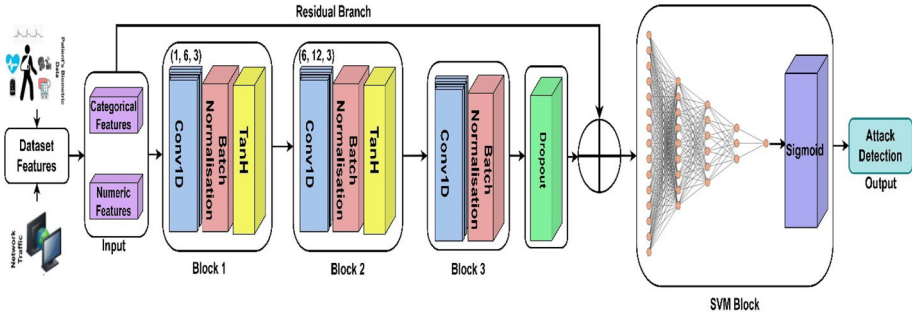


Fig. 4 Working flow of ConvSVM for detecting IoMT attack

3.4.3 CatEmb (categorical embedding)

Categorical entity embedding extracts the embedding layers of categorical variables from a neural network model and applies numeric vectors to represent the properties of the categorical values. It is usually used on categorical variables with high cardinalities. In this model, categorical features are separately handled by column embeddings. In contrast, the BN individually manages numerical features. Then, these features are concatenated together and fed to decider Block 1 containing the FC layer, BN, ReLU (Rectified Linear Unit) activation function, and dropout layer. In FC layers, the input gets multiplied by a weight matrix in an FC layer, and a bias vector gets added to it; BN helps to alleviate the model functionality by expediting the training process, employing more excellent learning rates, and eliminating internal covariate shifts; ReLU activation function fixes the vanishing gradients problem and does not allow activation of all the neurons at the same interval of time. Then, the dropout layer was used to drop noise or data value that was purposefully removed from a neural network to accelerate the processing of the model and restrain the model from overfitting. The output of block 1 is further sent to block 2, containing the same components. Then, the result of block 2 passed to the SVM block containing the FC layer and sigmoid activation function for the binary classification. Figure 5 depicts the working flow of categorical embedding for detecting IoMT attacks.

Table 5 shows the training parameters for all the proposed methods with learning rates

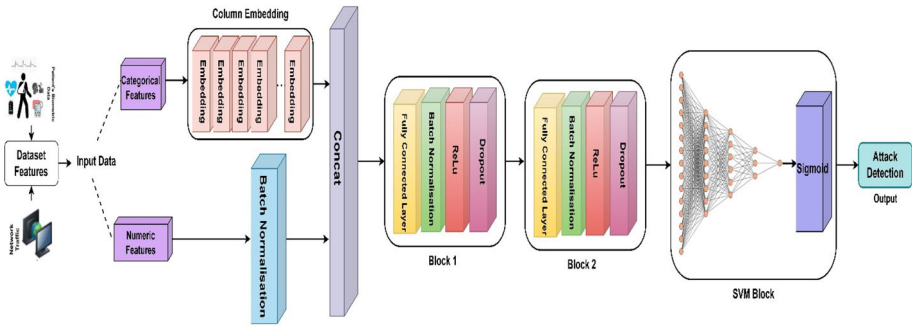


Fig. 5 Working flow of CatEmb for detecting IoMT attack

Table 5 Training parameters of the proposed models

Training parameters	DL models		
	LinSVM	ConvSVM	CatEmb
Initial state or attributes	in_features, hidden_size = 32, out_features = 1, n_hid_layer = 3	in_features = 40, out_features = 1	layers_dim = [64, 128] output_dim = 1 drop_prob = 0.5
Activation function	TanH	TanH	ReLU
Conv1D	–	Yes	–
BatchNorm1d	–	Yes	Yes
AdaptiveAvgPool1d	–	Yes	–
Batch size	64	64	64
Optimizer	Adam	Adam	Adam
Learning rate	0.0001	0.0001	0.0001

and batch sizes of 0.0001 and 64, respectively. In the further section, these models are also trained with different batch sizes like 128 and 256 to examine the methods' training loss and validation loss performance.

The CatEmb approach has superior performance compared to other DL models because of various factors inherent in its design which are explained as follows:

- Handling categorical variables: This model handles categorical variables with high cardinalities by using categorical entity embedding. This approach allows the model to present the properties of categorical values as numeric vectors and improves the model's ability to depict complex relationships within the data.
- Feature representation: This model employs column embeddings to manage categorical characteristics whereas BN handles numerical features individually. This distinct approach guarantees that the model accurately represents both categorical and numerical data.
- Neural network architecture: In this model, a neural network architecture consists of decider blocks (such as FC, BN, ReLU activation function, and dropout layers) and an SVM block. The decider blocks allow efficient feature extraction and learning. Furthermore, the model integrates dropout regularization to prevent overfitting and improve generalization.
- Efficient classification: The output from block 1 is further processed through block 2 before being transmitted to the SVM block for binary classification. This multi-layered approach allows the model to depict patterns and relationships in the data, which helps the model accurately classify the network intrusion.

These features help achieve superior results compared to other models and require less inference time to detect intrusion in the network.

4 Experimental setup and result discussion

This paper assesses ML and DL models using performance metrics, including accuracy, precision, recall, F1-score, specificity, and MCC. After that, this section highlights the working performance of the proposed model compared to the state of the art. Here, Fig. 6 repre-

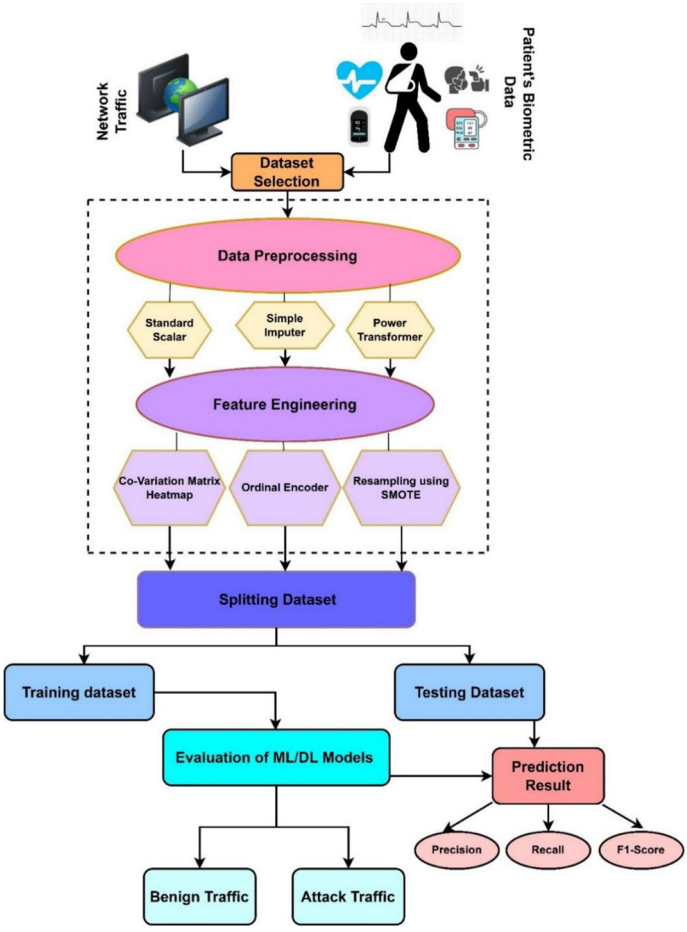


Fig. 6 Experiment outline to recognize IoMT attack

sents the experiment outline of this work. The WUSTL-EHMS-2020 dataset [11] has been selected as a network and patient biometric data combination. After that, data preprocessing techniques like standard scaler, simple imputer, and power transformer have been applied to clean and transform raw data into an appropriate format for the training process. Then, the feature engineering techniques, including co-variation matrix heatmap and ordinal encoder, were used to transform existing features or build new ones to enhance the performance of the models. The SMOTE function has been utilized for resampling the dataset to balance the classes. Further, the performance has been evaluated for 80–20% and 70–30% of dataset distributions for both ML and DL models to identify attacks in the IoMT environment.

To begin with, deep learning models, including LinSVM, ConvSVM, and CatEmb, are seamlessly integrated into the framework to enhance network-based attack detection in the IoMT). CatEmb is a novel embedding technique that employs DL techniques and recognizes intrusions using the WUSTL-EHMS-2020 real-time dataset. Additionally, the comprehensive evaluation using binary classification methods helps to separate benign network traffic from potential threats such as IoT data injection attacks and spoofing, guaranteeing useful

effectiveness in IoMT networks. Moreover, the memory consumption and inference time are optimized using extensive preprocessing and feature engineering techniques, such as systematic column reduction via covariance matrix analysis. The dataset is balanced through data augmentation methodologies, and feature segregation makes it easier to apply standard approaches such as ordinal encoders for categorical columns and standard scaler and power transformers for continuous columns. Finally, the optimal performance is achieved by seamlessly integrating with other components. This is confirmed by comprehensive evaluations on real-time benchmark dataset, which provide improved accuracy, precision, recall, and F1-score for attack detection. These results are achieved within a minimal inference time. The integration of DL models in IoMT environments highlights the effectiveness of these models in enhancing network security.

4.1 Hardware and software requirements

The experiment work has been implemented on Windows 10 (64-bit Operating System) with the configuration of 16 GB RAM, Intel(R) Core (TM) i5-7200U CPU @ 2.50 GHz. For training and testing the ML and DL models, python libraries like PyTorch (commonly utilized for developing deep neural networks) [56] and Scikit-learn [57] were used.

4.2 Healthcare testbed and dataset description

In [11], the author developed a testbed using a health monitoring sensor board. It gathered the information using various healthcare sensor devices affixed to the patient body. With a USB port's help, the panels connect to the window-based system, which uses C++ -based software for collecting the sensed data. The testbed combines components such as a sensor device that helps gather information related to patient health, an SDN controller, and the network gateway for envisaging the network flow traffic. The information generated using sensor and network traffic in the testbed gets utilized to identify the attacks and anomalies in the data. While developing the dataset, three attacks were considered, i.e., spoofing attacks, MITM attacks, and data injection. In the MITM attack, the intruder positioned himself as a router and received the message first. After altering or spoofing the message, reroute that file to the server. This results in an infringement of data integrity and confidentiality of the personal information in the network. In a spoofing attack, a program or a person pretends to be someone else to gain the victim's confidence so that the intruder can successfully enter the system, obtain information or money, and spread the virus. It results in an infringement of patient data protection and confidentiality. In a data alteration attack, the intruder manipulates or alters the information received at the intruder system from the gateway. The alteration could be made at random or following a rule. Then, the data is forwarded to the server. The modification by the intruder in the data causes a severe problem for the patient's health, as false diagnostics will be given to the individual depending on the alteration message.

The ARGUS Tool [58] creates this combined dataset of network traffic flow and biometric data features of patients. To attain the network traffic flow data, their traffic flow metrics and associated records were collected, along with the biometrics features containing diastolic blood pressure, temperature, respiration rate, systolic blood pressure, pulse rate, ECG ST segment information, heart rate, and peripheral oxygen saturation data. Table 6 shows different features of the dataset. All-inclusive, the dataset includes 44 features, out of which 35 features belong to the network traffic flow category. The dataset output is categorized as

Table 6 Description of ML features in the dataset

Metric	Type	Features description	Metric datatypes
Flgs	Flow Metric	Flags in the network	Int
SIntPkt	Flow Metric	Source inter-packet count	Float
DIntPkt	Flow Metric	Destination inter-packet count	Float
SrcBytes	Flow Metric	Source bytes	Float
DstBytes	Flow Metric	Destination bytes	Float
SrcLoad	Flow Metric	Source load	Float
DstLoad	Flow Metric	Destination load	Float
SreGap	Flow Metric	Source missing bytes	Float
DstGap	Flow Metric	Destination missing bytes	Float
sMinPktSz	Flow Metric	The minimum packet size of traffic sent from the source	Float
dMinPktSz	Flow Metric	The minimum packet size of traffic sent from the destination	Float
sMaxPktSz	Flow Metric	The maximum packet size of traffic sent from the source	Float
dMaxPktSz	Flow Metric	The maximum packet size of traffic sent from the destination	Float
SIntPktAct	Flow Metric	Source active inter-packet arrival time	Float
DIntPktAct	Flow Metric	Destination active inter-packet arrival time	Float
SrcMac	Flow Metric	Source Mac Address	Int
DtMac	Flow Metric	Destination Mac Address	Int
Loss	Flow Metric	Dropped packets or Retransmitted	Float
pLoss	Flow Metric	Percentage of dropped packets or retransmitted	Float
SrcJitter	Flow Metric	Source jitter	Float
DstJitter	Flow Metric	Destination jitter	Float
Trans	Flow Metric	Cumulative packets count	Float
Dur	Flow Metric	Total Duration	Float
TotPkts	Flow Metric	Total number of packets count	Float
TotBytes	Flow Metric	Total number of packets bytes	Float
pSrcLoss	Flow Metric	Percentage of source dropped packets count or retransmitted packets	Float
pDstLoss	Flow Metric	Percentage of destination dropped packets count or retransmitted packets	Float
Rate	Flow Metric	Number of packet counts per second	Float
Load	Flow Metric	Load	Float
Heart_Rate	Biometric	Heart rate	Float
SpO2	Biometric	Peripheral oxygen saturation	Float
Temp	Biometric	Temperature	Float

Table 6 (continued)

Metric	Type	Features description	Metric datatypes
DIA	Biometric	Diastolic blood pressure	Float
Resp_Rate	Biometric	Respiration rate	Float
SYS	Biometric	Systolic blood pressure	Float
Pulse_Rate	Biometric	Pulse rate	Float
ST	Biometric	ECG ST segment	Float

regular traffic or attack traffic data. "0" represents the normal traffic, while "1" illustrates attack traffic data.

4.3 Performance evaluation metrics

The statistical parameters are applied to determine the working performance of the proposed models. To assess the performance of IoMT attack classification, we used four well-known metrics, which are explained as follows:

Using the above metrics shown in Table 7, the working performance of the ML and DL models gets assessed in terms of different standard metrics as discussed below:

Accuracy is elucidated as the ratio of the correct traffic categorization to the overall prediction of the test data for network traffic and is measured using Eq. (12). Precision is elucidated as the ratio of the correct categorization of attack traffic to the total of the correct and incorrect categorization of attack traffic samples in the dataset. The model functions well if the accuracy is high and is measured using Eq. (13). The recall/sensitivity is elucidated as the ratio of the correctly classified attack network traffic to the total of precisely classified attack network traffic and incorrectly classified attack network traffic in the dataset and is measured using Eq. (14). F1-score amalgamates the precision and recall of a classifier into a single statistic by calculating their harmonic mean as expressed in Eq. (15). Specificity is a performance evaluation standard that assesses a model's accuracy in detecting adverse events. It is especially crucial when accurately recognizing true negatives is essential, like in fraud detection or medical tests. It is measured using Eq. (16). MCC (Matthews correlation

Table 7 Components of confusion matrix

Components of confusion matrix	Description
True positive (TP)	In this case, both the forecasted and actual outcomes of the data point are valid. It is appropriately categorized as an attack sample
True negative (TN)	In this case, the data point's forecasted and actual outcomes are false. Therefore, it is appropriately categorized as a standard sample
False positive (FP)	In this case, the forecasted outcome of the data point is precise, and the actual result of the data point is false. It designated a standard data set as an attack sample
False negative (FN)	In this case, the forecasted outcome of the data point is false, whereas the actual result is valid. Therefore, it was restricted as a standard sample despite being an attack sample

coefficient) is a statistical tool is used to evaluate the model to measure the diversion of the actual values from the predicted values, similar to the chi-square measure in reference to a 2×2 contingency table. It is measured using Eq. (17). NVP (negative predictive value) is a statistical index used to analyze diagnostic accuracy evaluations and medical tests. It shows the likelihood that a person who tests negative on a diagnostic or screening test does not have the disorder or disease the examination intends to diagnose. The accuracy and significance of a negative test result are more likely to be determined by NPV. It is measured using Eq. (18). FPR (false positive rate) measure shows the module's outcomes that incorrectly predict the positive class as expressed in Eq. (19). FNR (false negative rate) measure shows those outcomes of the model which incorrectly predict the negative class as described in Eq. (20). FDR (false discovery rate) calculates the rate at which one mistakenly believes a null hypothesis is untrue. It is expressed using Eq. (21). FOR (false omission rate) is a statistical indicator called false omission rate (FOR), which is used in binary classification to evaluate the percentage of actual positive events that a model incorrectly predicts as negative. It is expressed using Eq. (22). ROC AUC (receiver operating characteristic area under the curve): The capacity of a model to discern between positive and negative classes across various classification thresholds is assessed using the ROC_AUC performance metric for binary classification models.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{13}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{14}$$

$$F1 - \text{Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \tag{15}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{16}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{17}$$

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{18}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{19}$$

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{20}$$

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}} \tag{21}$$

$$\text{FOR} = \frac{\text{FN}}{\text{FN} + \text{TN}} \tag{22}$$

4.4 Result analysis of ML models

This section highlights the evaluation performance of ML models like SVM, DT, LR, KNN, GradBoost, RF, and XGB using tenfold cross-validation on the WUSTL-EHMS-2020 dataset.

Table 8 Performance of ML models using tenfold cross-validation on the WUSTL-EHMS-2020 dataset

Folds	SVM				DT				LR				KNN			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
0	0.9996	0.9993	1.0	0.9996	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9984	0.9972	0.9996	0.9984
1	0.9998	0.9996	1.0	0.9998	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9987	0.9978	0.9996	0.9987
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9980	0.9964	0.9996	0.9980
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9984	0.9968	1.0	0.9984
4	0.9996	0.9993	1.0	0.9996	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9987	0.9982	0.9993	0.9987
5	0.9998	0.9996	1.0	0.9998	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9985	0.9978	0.9992	0.9985
6	0.9998	0.9996	1.0	0.9998	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9984	0.9978	0.9989	0.9984
7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9984	0.9972	0.9996	0.9984
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9985	0.9979	0.9993	0.9986
9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9991	0.9989	0.9993	0.9991
Average Score	0.9998	0.9997	1.0	0.9998	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9985	0.9976	0.9994	0.9985

Folds	Grad_Boost				RF				XGB			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
3	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
4	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
6	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
7	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 8 (continued)

Folds	Grad_Boost				RF				XGB			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
8	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Average Score	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

SVM and KNN achieved an average accuracy of 0.9998 and 0.9985, respectively, whereas other models attained 100% accuracy, as depicted in Table 8.

This section evaluates ML models by considering different performance metrics parameters like accuracy, precision, recall, F1-score, specificity, MCC, NPV, FPR, FNR, FDR, FOR, and ROC_AUC, as shown in Table 9. The inference time has also been considered here to check which model quickly recognizes attacks in the IoMT network with better accuracy. The performance has been evaluated for 70–30% and 80–20% dataset distribution. Here, SVM and KNN achieved 99.97% and 99.99% accuracy, respectively, whereas other trained models attained 100% accuracy. LR model takes less inference time, whereas RF takes the highest time among different models to identify an attack in the IoMT environment.

4.5 Result analysis of proposed DL models

This section evaluates DL models by considering different performance metrics parameters like accuracy, precision, recall, F1-score, specificity, MCC, NPV, FPR, FNR, and ROC_AUC, as shown in Table 10. The inference time has also been considered here to check which model quickly recognizes attacks in the IoMT network with better accuracy. The performance has been evaluated for 70–30% and 80–20% dataset distribution, as demonstrated in Fig. 7. Here, LinSVM achieved 99.96% accuracy, whereas other trained models like ConvSVM and CatEmb attained 100% accuracy. Among two models, i.e., LinSVM and ConvSVM, LinSVM model takes less Inference time, whereas ConvSVM takes the highest time to identify attacks in the IoMT environment, as depicted in Fig. 7 (i).

Based on the performance analysis, both LinSVM and ConvSVM models demonstrate optimal results on each metrics as depicted in Table 10. However, LinSVM model takes less time (0.1629 s) as compared to ConvSVM model (0.3116 s) for detecting intrusion in the network. In IoMT environment, timely detection of network intrusion is critical for ensuring safety and data security of patients. Therefore, based on the above-achieved results, LinSVM model appears to be the more suitable choice for deployment in resource-constrained IoMT environment.

4.6 Performance comparison of proposed models with existing ML and DL approaches

This section compares proposed ML and DL models with the existing methods, as shown in Table 11. Here, different existing methods have been proposed by the experts to identify attacks in the IoMT environment using IDS frameworks, which are discussed as follows: In [59], Dina et al. proposed feed-forward neural network model (FFNN) with the focal loss method to detect intrusion in the network and achieved 93.26% of accuracy. The author applied the focal loss function to remove the data imbalance challenge. In [60], Chaganti et al. developed a PSO-DNN (particle swarm optimization deep neural network) model to recognize intrusion in the network and achieved 96% accuracy. Furthermore, Wagan et al. [61] proposed a model combination of Bi-LSTM and fuzzy mean clustering to collect sensory medical data safely and recognize attack patterns in the IoMT environment. The proposed model attained an accuracy of 92.63%. In this work, the author mainly focuses on enhancing the performance of the IDs and improving the detection rate (DR), which is underutilized as the author hasn't analyzed both delay and DR simultaneously. The results achieved by the proposed models show that our models are superior to the existing approaches in terms of performance metrics like accuracy, precision, recall, F1-score, and MCC, as shown in Fig. 8.

Table 9 The working performance analysis of ML models

Metrics	ML Techniques													
	SVM		DT		LR		KNN		GBM		RF		XGB	
	70–30%	80–20%	70–30%	80–20%	70–30%	80–20%	70–30%	80–20%	70–30%	80–20%	70–30%	80–20%	70–30%	80–20%
TP	2859.3	2859.4	2859.8	2859.0	2859.8	2859.0	2854.5	2856.4	2859.8	2859.0	2859.8	2859.0	2859.8	2859.0
TN	2847.0	2847.6	2848.2	2849.0	2848.2	2849.0	2836.2	2842.5	2848.2	2849.0	2848.2	2849.0	2848.2	2849.0
FP	1.2	0.6	0.0	0.0	0.0	0.0	12.0	6.5	0.0	0.0	0.0	0.0	0.0	0.0
FN	0.5	0.4	0.0	0.0	0.0	0.0	5.3	2.6	0.0	0.0	0.0	0.0	0.0	0.0
Accuracy	0.9998	0.9997	1.0	1.0	1.0	1.0	0.9998	0.9999	1.0	1.0	1.0	1.0	1.0	1.0
Precision	0.9995	0.9997	1.0	1.0	1.0	1.0	0.9958	0.9977	1.0	1.0	1.0	1.0	1.0	1.0
Recall	0.9998	0.9998	1.0	1.0	1.0	1.0	0.9981	0.9990	1.0	1.0	1.0	1.0	1.0	1.0
F-score	0.9997	0.9998	1.0	1.0	1.0	1.0	0.9969	0.9984	1.0	1.0	1.0	1.0	1.0	1.0
Specificity	0.9995	0.9997	1.0	1.0	1.0	1.0	0.9957	0.9977	1.0	1.0	1.0	1.0	1.0	1.0
MCC	0.9994	0.9996	1.0	1.0	1.0	1.0	0.9939	0.9968	1.0	1.0	1.0	1.0	1.0	1.0
NVP	0.9998	0.9998	1.0	1.0	1.0	1.0	0.9981	0.9990	1.0	1.0	1.0	1.0	1.0	1.0
FPR	0.0004	0.0002	0.0	0.0	0.0	0.0	0.0042	0.0022	0.0	0.0	0.0	0.0	0.0	0.0
FNR	0.0001	0.0001	0.0	0.0	0.0	0.0	0.0018	0.0009	0.0	0.0	0.0	0.0	0.0	0.0
FDR	0.0005	0.0003	0	0	0	0	0.0042	0.0023	0	0	0	0	0	0
FOR	0.0002	0.0002	0	0	0	0	0.0019	0.0010	0	0	0	0	0	0
ROC_AUC	0.9997	0.9998	1.0	1.0	1.0	1.0	0.9969	0.9984	1.0	1.0	1.0	1.0	1.0	1.0
Inference Time (Sec)	0.0645	0.05377	0.0154	0.02022	0.0143	0.01181	0.02237	0.03951	0.02492	0.01978	0.09705	0.05927	0.02477	0.01832

Table 10 The working performance analysis of DL models

Metrics	DL models					
	LinSVM		ConvSVM		CatEmb	
	70–30%	80–20%	70–30%	80–20%	70–30%	80–20%
TP	4212	2791	4211	2791	4211	2791
TN	4345	2903	4347	2905	4349	2905
FP	1	0	2	0	0	0
FN	2	2	0	0	0	0
Accuracy	0.9996	0.9996	0.9997	1.0	1.0	1.0
Precision	0.9997	1.0	0.9995	1.0	1.0	1.0
Recall	0.9995	0.9992	1.0	1.0	1.0	1.0
F-score	0.9996	0.9996	0.9997	1.0	1.0	1.0
Specificity	0.9997	1.0	0.9995	1.0	1.0	1.0
MCC	0.9992	0.9992	0.9995	1.0	1.0	1.0
NVP	0.9995	0.9993	1.0	1.0	1.0	1.0
FPR	0.0002	0.0	0.0004	0.0	0.0	0.0
FNR	0.0004	0.0003	0.0	0.0	0.0	0.0
FDR	0.0003	0	0.0005	0	0	0
FOR	0.0005	0.0007	0	0	0	0
ROC_AUC	0.9996	0.9996	0.9997	1.0	1.0	1.0
Inference Time (Sec.)	0.1029	0.1629	1.1548	0.3116	0.0335	0.0396

The proposed models took minimum intrusion detection time as compared to other ML and DL approaches.

4.7 Training and validation loss for DL models: *an ablation study*

The continuous monitoring of training and validation losses guarantees that the model is learning efficiently without overfitting and aids in the training process. Therefore, this section highlights these losses for different numbers of epochs that have been calculated along with training and testing accuracy and Inference time taken by the models to identify the attack in the network.

4.7.1 Training and validation loss for dl models after ($n \times 50$) Epochs

In this section, training and validation loss has been evaluated for LinSVM, ConvSVM, and CatEmb models after every ($n \times 50$) Epochs, where $n = 1,2 \ldots .., 4$ as shown in Table 12. To begin with model performance for 50 epochs, the CatEmb model performed better by attaining the highest accuracy among the other two models with 0.0285 training loss and 0.1562 validation loss. In addition to performance for 100 epochs, LinSVM has the lowest training and validation loss compared to the other models, with an accuracy of 99.98%. Furthermore, for 150 epochs, the ConvSVM model performed better by achieving the highest accuracy rate along with 0.0030 training loss and 0.0002 validation loss. In contrast, LinSVM models

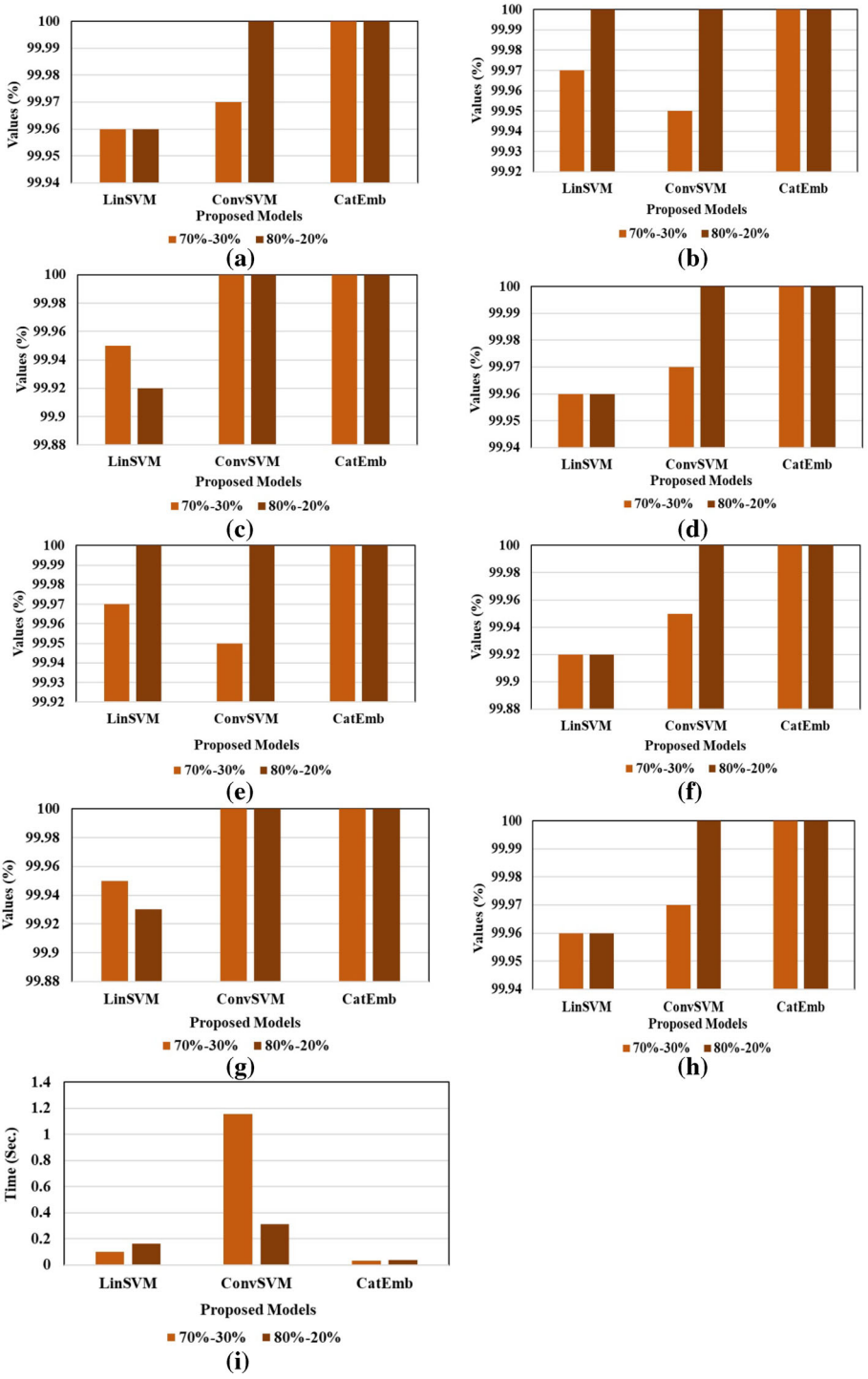


Fig. 7 The working performance analysis of DL models **a** Accuracy, **b** precision, **c** recall, **d** F1-score, **e** Specificity, **f** MCC, **g** NVP, **h** ROC_AUC, **i** Inference time

Table 11 Performance comparison of proposed models with ML and DL-based existing methods

Author	Proposed method	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	MCC (%)	Inference time (Sec.)
Saif et al. [28]	DHOA-ANFIS	95.88	94.44	95.30	–	–	–
Saif et al. [40]	PCA-RF	99	100	100	100	–	49.23
Saif et al. [41]	GA-DT	99.88	–	–	–	–	–
Dina et al. [56]	FFNN-Focal Loss	93.26	95.24	73.69	80.11	65.48	–
Chaganti et al. [57]	PSO-DNN	96	96	96	95	–	–
Wagan et al. [58]	Fuzzy C-Means	92.63	92.05	96.14	96.9	–	–
Proposed methods	LinSVM	99.96	100	99.92	99.96	99.92	0.16297
	ConvSVM	100	100	100	100	100	0.31163
	CatEmb	100	100	100	100	100	0.0396

attained a training loss of 0.0001 and a validation loss of 0.0062, and CatEMB achieved the highest training and validation loss compared to other models. Finally, for 200 epochs, the ConvSVM model performed better by attaining a minimum training loss of 0.0013 and validation loss of 0.0004 with the highest training and testing accuracy compared to the other models.

In our research, we have noted improved scalability features within our proposed models by incorporating diverse and dependable data sources from various IoT devices into the training dataset. This integration ensures a high level of coherence and a wider range of data. Significantly, when the epoch count was increased from 50 to 200 ($(n \times 50)$ epochs, where $n = 1,2\ldots..,4$), the accuracy of our suggested model demonstrated minor fluctuations, highlighting its capacity to scale well. We used real-time dataset for conducting the training and testing of proposed model. Figure 9 shows the scalability analysis of our proposed models.

4.7.2 Training and validation plots for different batch sizes

This section demonstrates the training and validation loss plots on different batch sizes, including 64, 128, and 256. In Fig. 10a, c, and e, the training loss is measured on the vertical axis (Y), while the number of epochs is displayed on the horizontal axis (X). It is calculated for the number of epoch counts ranging from 1 to 200 for different batch sizes like 64, 128, and 256 using proposed DL models, namely LinSVM, ConvSVM, and CatEmb. It is observed from Fig. 10a, c, and e that LinSVM and ConvSVM converge to 0 after 50 epochs in all the batch sizes. Figure 10b, d, and f depicts the validation loss that is measured on the vertical axis (Y), while the number of epochs is displayed on the horizontal axis (X). The validation loss is calculated for the number of epoch counts ranging from 1 to 200 for different batch sizes using proposed DL models. It is observed from Fig. 10b, d, and f that after 50 epochs,

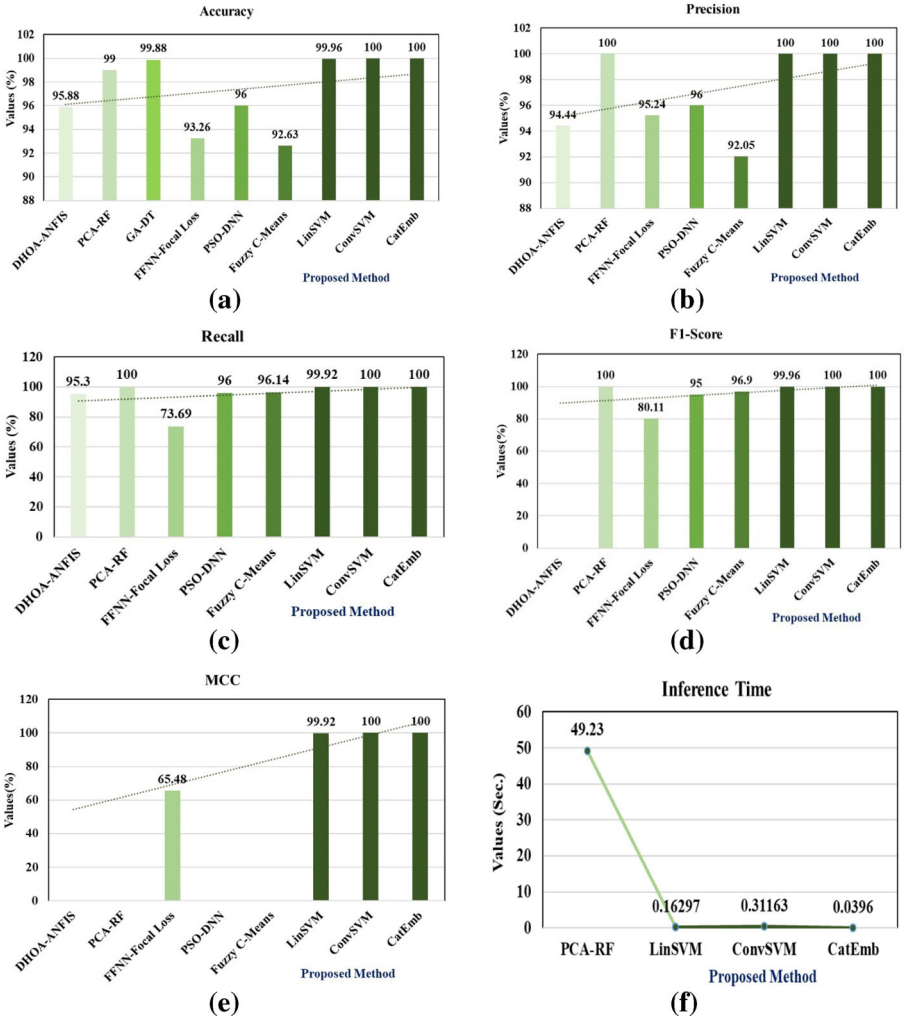


Fig. 8 Performance comparison of proposed models with DL-based existing methods **a** accuracy, **b** precision, **c** recall, **d** F1-score, **e** MCC, **f** inference time

ConvSVM converges to 0 rapidly as compared to the other two models for batch size 64. For 128 and 256 batch sizes, LinSVM and ConvSVM converge to 0 after 50 epochs.

In Table 13, the proposed DL models have been evaluated for different batch sizes like 64, 128, and 256 with a learning rate of 0.0001. To begin with, 64-batch size, LinSVM, ConvSVM, and CatEmb attained an accuracy of 99.96%, 100%, and 100%, respectively; in addition to performance evaluation for 128 batch size, all the proposed dl models achieved maximum accuracy. Furthermore, for 256-batch size, LinSVM attained an accuracy of 99.94%, whereas ConvSVM and CatEmb achieved 100% accuracy. Figure 11 depicts the performance of the proposed model for batch sizes 64, 128, and 256. For batch size 256, models take less inference time than batch sizes 64 and 128, as shown in Fig. 11d.

Table 12 Training and validation loss for ($n \times 50$) epochs

No. of epochs	Metrics	DL models		
		LinSVM	ConvSVM	CatEmb
50 epochs	TA	0.9931	0.9826	0.9984
	TA-IT	0.0466	0.4280	0.0512
	TL	0.244	0.0559	0.0285
	VL	0.034	0.0767	0.1562
	TA*	0.9992	0.9949	1.0
	TA*-IT	0.0093	0.1473	0.0016
100 epochs	TA	0.9998	0.9992	0.9984
	TA-IT	0.0600	0.3127	0.0248
	TL	0.0006	0.0151	0.0461
	VL	0.0006	0.0191	0.1562
	TA*	0.9998	0.9992	1.0
	TA*-IT	0.01692	0.0485	0.0080
150 epochs	TA	0.9984	0.9998	0.9984
	TA-IT	0.08653	0.3681	0.0229
	TL	0.0001	0.0030	0.0076
	VL	0.0062	0.0002	0.1562
	TA*	0.9994	1.0	1.0
	TA*-IT	0.01752	0.0730	0.0059
200 epochs	TA	0.9975	0.9996	0.9984
	TA-IT	0.0717	0.3438	0.0209
	TL	0.0260	0.0013	0.0230
	VL	0.0816	0.0004	0.1562
	TA*	0.9992	1.0	1.0
	TA*-IT	0.0260	0.1627	0.0029

TL Training loss, *VL* Validation loss, *TA* Training accuracy, *TA-IT* Training inference time, *TA** Testing accuracy, *TA*-IT* Testing inference time

4.8 Performance evaluation of proposed approach with state-of-the-art

This section highlights the comparison of the proposed model with the state-of-the-art model. Our models achieved the highest performance compared to the other existing work, as shown in Table 14. Researchers have applied ML and DL approaches for designing an improved IDS, which are discussed and compared with the proposed models in this section. In [11], Hady et al. developed an upgraded healthcare testbed for monitoring the patients and used it to gather biometrics data of the patient along with network flow metrics. In this work, the author built a dataset that combines the feature of both biometric data of the patients and network flow metrics and then applied different ML approaches, including ANN, SVM, RF, and KNN, and attained an accuracy of 88.75%, 92.44%, 92.06%, and 92.27%, respectively. ANN achieved the lowest accuracy as compared to other models. In [62], Salman et al. recommend that any measure used to assess an ML model created for a security application

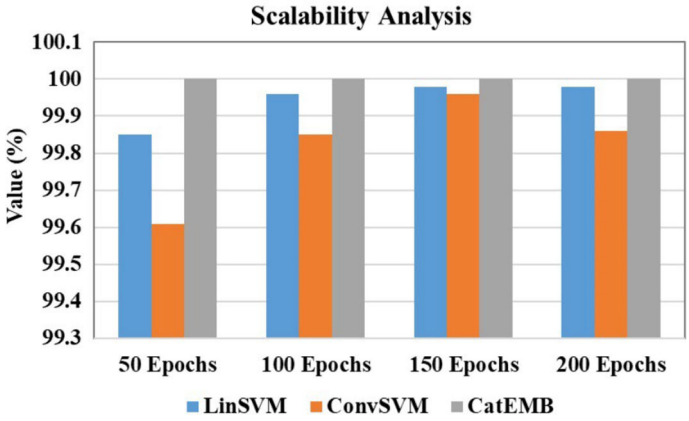


Fig. 9 Scalability analysis of proposed models

meet these two criteria: (1) the errors committed in incorrectly identifying attacks and benign activity shouldn't receive the same treatment, and (2) when a sample is accurately classified in a class, it shouldn't be mistakenly misclassified in the same class. Therefore, the authors suggested a safety score metric based on the cost related to the model results and fulfilling both criteria. Here, the author used three ml models, including LR, KNN, and RF, which achieved an accuracy of 87.3%, 88.5%, and 93%, respectively. In [63], Gupta et al. proposed a tree classifier to recognize intrusion in the network and attained an accuracy of 94.23%. In [64], Kilincer et al. proposed the XGB and LR model with tenfold cross-validation on four different datasets: the ICU dataset, ECU-IoHT, WUSTL-EHMS, and TON-IoT. For the WUSTL-EHMS dataset, the proposed model attained 96.2% of accuracy. For encoding IoMT data, Dhanya and Chitra et al. [65] proposed an autoencoder-based DL framework that fed the encoded features to the optimized XGB classifier to detect the malware in the network and achieved an accuracy of 98.98%. The XGB classifier was optimized using a genetic algorithm. In [66], Dhanya and Chitra proposed an optimized XGB to detect malware in the IoMT network. For XGB hyperparameter optimization, the differential evolution (DE) method was used by the author, who achieved an accuracy of 97.39%. Compared to the state of the art, our proposed models achieved 100% accuracy, which is superior to the existing work performance, as depicted in Fig. 12.

5 Conclusion and future scope

IoMT services are in more demand than ever before as the number of users has been increasing rapidly. This calls for a robust security model to thwart any unwanted activities. An attack detection method is also required in the security architecture for distributive networks to determine whether an intrusion has affected the network's data. The goal of this research is threefold: firstly, we suggest the linear support vector machine (LinSVM) model; secondly, the convolutional support vector machine (ConvSVM) model; and finally, the categorical embedding (CatEmb) model which have been proposed to overcome the issue of security in a network. This article offers the CatEmb model as the first effort to use a DL-based embedding approach to recognize intrusion in the IoMT environment. Compared to the other models,

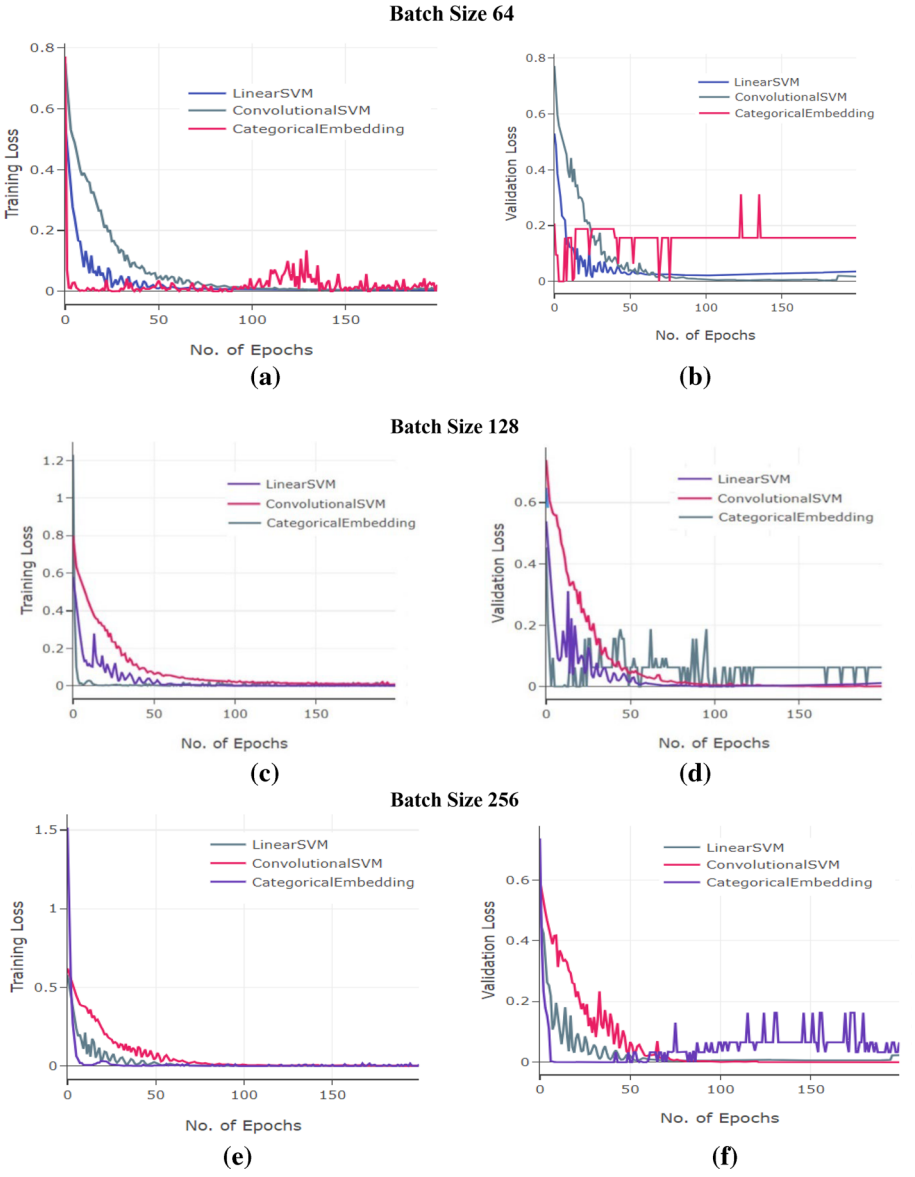


Fig. 10 Training and validation loss for DL models for batch sizes 64, 128, and 256. Batch size 64 **a** training loss for DL models, **b** validation loss for DL models, batch size 128 **c** Training Loss for DL models, **d** validation loss for DL models. Batch size 256 **e** training loss for DL models. **f** Validation loss for DL models

CatEmb takes less inference time in detecting attacks in the network. With a 100% accuracy, detection rate, and F1-score, our developed models exceed the state of the art in detecting network intrusions by utilizing the combined features of the network traffic flow metric and the patient's recent biometric dataset. In addition, an in-depth evaluation of different ML models, including SVM, LR, DT, KNN, GradBoost, RF, and XGBoost, has been explored

Table 13 Performance analysis of DL models for different batch sizes

Batch size	Model	TL	VL	TA	TA*	*P*	*R*	F1	IT (Sec.)
64	LinSVM	0.0260	0.0816	0.9978	0.9996	1.0	0.9992	0.9996	0.162
	ConvSVM	0.0013	0.0004	0.9998	1.0	1.0	1.0	1.0	0.31163
	CatEmb	0.0230	0.1562	0.9984	1.0	1.0	1.0	1.0	0.0396
128	LinSVM	0.0001	0.0114	0.9978	1.0	1.0	1.0	1.0	0.08987
	ConvSVM	0.0073	0.0011	0.9996	1.0	1.0	1.0	1.0	0.73434
	CatEmb	0.0025	0.0625	0.9993	1.0	1.0	1.0	1.0	0.01614
256	LinSVM	0.0001	0.0226	0.9984	0.9994	1.0	0.9994	0.9989	0.04434
	ConvSVM	0.0021	0.0605	1.0	1.0	1.0	1.0	1.0	0.27644
	CatEmb	0.0078	0.0651	0.9993	1.0	1.0	1.0	1.0	0.01838

TL Training loss, *VL* Validation loss, *TA* Training accuracy, *TA** Testing accuracy, *P* Precision, *R* Recall, *F1* F1-score, *IT* Inference time in seconds

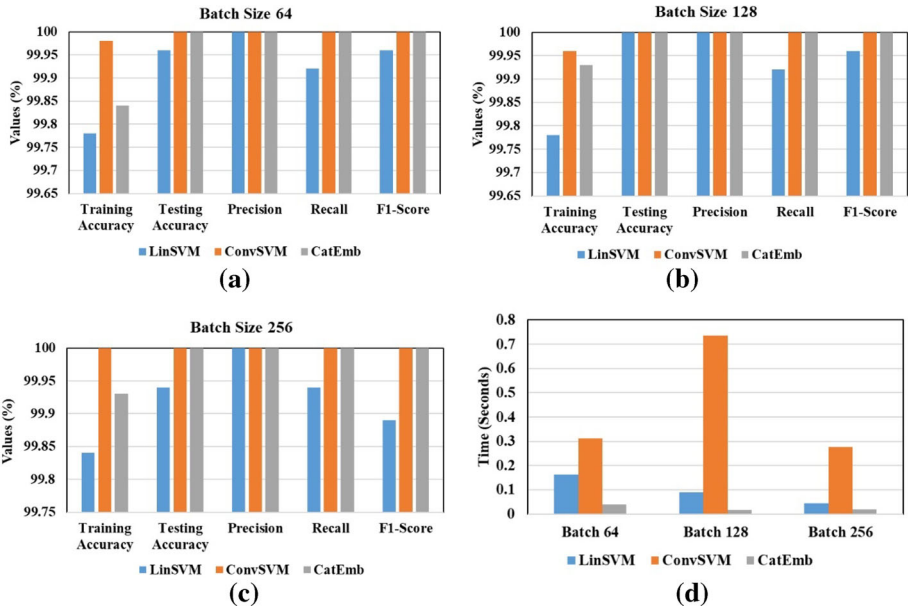


Fig. 11 Performance analysis of DL models for different batch sizes. **a** Batch size 64, **b** batch size 128, **c** batch size 256, **d** inference time comparison

for intrusion detection in the IoMT ecosystem. Data preprocessing and feature engineering techniques play a vital role in achieving such an outstanding result by handling categorical data, imbalanced, noisy data, and dropping irrelevant columns, which may not significantly affect the detecting process. This helps build an effective model that takes less memory utilization and inference time to detect attacks in the network. Furthermore, the performance of DL models has been evaluated for the batch sizes 64, 128, and 256 to calculate the training

Table 14 Performance evaluation of proposed models with state of the art

Dataset	Year	Authors	Methodology	*A* (%)	*P* (%)	*R* (%)	F1 (%)	ROC_AUC (%)	Description	Observations
WUSTL-EHMS-2020	2020	Hady et al. [11]	SVM	92.4	–	–	–	82.37	The author merged the features of network flow traffic and patients' biometric features data	Needing to enhance the system's performance, and in experimental work, the author only considered accuracy and ROU_AUC metric in the position, which might not deliver a comprehensive grasp of the model's capabilities
			RF	92.13	–	–	–	91.45		
			ANN	90.04	–	–	–	93.42		
			KNN	92	–	–	–	87.48		
	2020	Salman et al. [62]	LR	87.3	–	–	50.9	–	The authors suggested a Safety score metric based on the cost related to the model results	Only certain dataset features were considered by the author in this work, and accuracy and F1-score were calculated, which may not have given readers a complete understanding of the model's potential
			KNN	88.5	–	–	65.5	–		
			RF	93	–	–	71.7	–		

Table 14 (continued)

Dataset	Year	Authors	Methodology	*A* (%)	*P* (%)	*R* (%)	F1 (%)	ROC_AUC (%)	Description	Observations
			CNN-Focal	93.08	94.23	73.38	79.63	–		
	2022	Gupta et al. [63]	Tree Classifier	94.23	93.8	93.72	93.8	90.68	The author proposed a tree classifier to recognize intrusion in the IoMT network and enhance the performance compared to [8, 50]	Unrealistic datasets were produced via data augmentation. The percentage of attack traffic in the network is relatively low
	2022	Chaganti et al. [60]	PSO-DNN	96	96	96	95	–	The author proposed a PSO-DNN-based framework for detecting intrusion in the network and enhanced the working performance of the models compared to [11, 56, 59, and 60]. For data preprocessing, standard scalers and simple imputers were used	Performance can be further enhanced using feature engineering techniques
			LR	95.2	95	95	94	–		
			CNN	95.6	96	95	95	–		
			LSTM	95.2	96	95	95	–		

Table 14 (continued)

Dataset	Year	Authors	Methodology	*A* (%)	*P* (%)	*R* (%)	F1 (%)	ROC_AUC (%)	Description	Observations
	2022	Kilincer et al. [64]	XGB	96.2%	96.19	96.16	96.17	–	In this paper, the author used two ML techniques, including XGBoost and LR, based on recursive feature elimination (RFE) and then applied them to four different datasets. RFE is utilized in this work so that the working functionality of chosen features can be easily managed and achieve high accuracy compared to filter methods	In this work, the author only considered limited data features from the dataset. The performance can be improved using deep learning models
	2023	Dina et al. [59]	FFNN-focal loss	93.26	95.24	73.69	80.11	–	The author used the focal loss technique in this paper to remove the data imbalance challenge	The working performance of the model can be upgraded by applying proper feature engineering. In this work, the author only used the Focal Loss method, a traditional approach, to mitigate the data imbalance issue
			CNN-focal	93.08	94.23	73.38	79.63	–		

Table 14 (continued)

Dataset	Year	Authors	Methodology	*A* (%)	*P* (%)	*R* (%)	F1 (%)	ROC_AUC (%)	Description	Observations
	2023	Wagan et al. [61]	Fuzzy C-means	92.63	92.05	96.14	96.9	–	The author proposed a model combination of Bi-LSTM and fuzzy mean clustering to safely collect sensory and medical data and recognize attack patterns in the IoMT environment	In this work, the author mainly focuses on enhancing the performance of the IDs and improving DR, which is underutilized as the author hasn't analyzed both delay and DR simultaneously
	2023	Dhanya & Chitra [65]	Optimized XGB classifier	98.98	93.02	98.97	95.91	99.64	This paper proposed a lightweight autoencoder-based DL framework for detecting malware in the IoMT environment with less computation time	The author does not address potential limitations or challenges associated with the proposed framework, such as its scalability, robustness to different types of malware, or generalizability to diverse IoMT environments

Table 14 (continued)

Dataset	Year	Authors	Methodology	*A* (%)	*P* (%)	*R* (%)	F1 (%)	ROC_AUC (%)	Description	Observations
	2023	Dhanya & Chitra [66]	Differential Evolution (DE)-XGB	97.39	–	–	–	–	The author suggested an extreme gradient boosting technique to classify attacks in the IoMT environment. DE, an intelligent evolutionary approach, is used to optimize the hyperparameters of the XGB algorithm	The author only considered the accuracy metrics in the work, which might not deliver a comprehensive grasp of the model's capabilities
	Proposed method		LinSVM	99.96	100	99.92	99.96	99.96	Proposed DL models that attained the highest performance as compared to the state-of-the-art to detect different attacks in the IoMT environment in less inference time, and minimum training and validation loss	–
			ConvSVM	100	100	100	100	100		
			CatEmb	100	100	100	100	100		

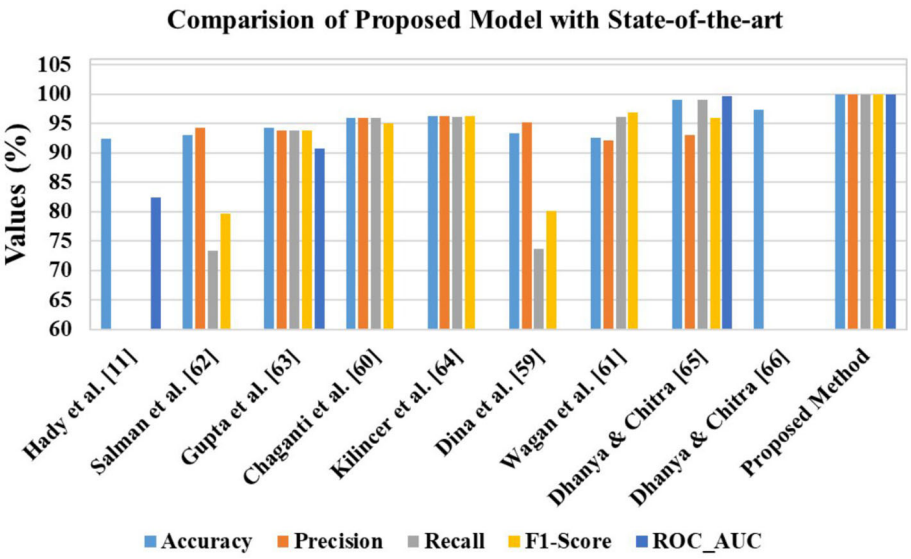


Fig. 12 Comparison of proposed model with state of the art

and validation loss, which are less than 0.0260 and 0.0816, respectively. In future, we will integrate Blockchain technology with the deep learning approach and test it on other existing IDS datasets and real-world testbeds comprised of patient biometrics and network features to detect intrusion by considering more emerging adversarial attacks in the network.

Authors' contributions All the authors contributed equally.

Data and materials availability Data will be made available on request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This study uses publicly available data or data from published sources; therefore, no subject testing or data collection procedure was considered for this study.

References

1. Park C, Lee J, Kim Y, Park J-G, Kim H, Hong D (2023) An enhanced AI-based network intrusion detection system using generative adversarial networks. IEEE Internet Things J 10(3):2330–2345. https://doi.org/10.1109/jiot.2022.3211346
2. Li J, Tong X, Liu J, Cheng L (2023) An efficient federated learning system for network intrusion detection. IEEE Syst J 17(2):2455–2464. https://doi.org/10.1109/jsyst.2023.3236995
3. Wang K, Zhang A, Sun H, Wang B (2022) Analysis of recent deep-learning-based intrusion detection methods for in-vehicle network. IEEE Trans Intell Transp Syst. https://doi.org/10.1109/tits.2022.3222486
4. Khoa TV, Hoang DT, Trung NL, Nguyen CT, Quynh TT, Nguyen DN, Ha NV, Dutkiewicz E (2023) Deep transfer learning: a novel collaborative learning model for Cyberattack detection systems in IOT Networks. IEEE Internet Things J 10(10):8578–8589. https://doi.org/10.1109/jiot.2022.3202029

5. Nandanwar H, Katarya R (2024) Deep learning enabled intrusion detection system for industrial IOT environment. Expert Syst Appl 249:123808. https://doi.org/10.1016/j.eswa.2024.123808
6. Sharma B, Sharma L, Lal C, Roy S (2023) Anomaly based network intrusion detection for IOT attacks using deep learning technique. Comput Electr Eng 107:108626. https://doi.org/10.1016/j.compeleceng.2023.108626
7. Aljuhani A, Alamri A, Kumar P, Jolfaei A (2024) An intelligent and explainable SAAS-based Intrusion Detection System for resource-constrained IoMT. IEEE Internet Things J. https://doi.org/10.1109/jiot.2023.3327024
8. Yan F, Zhang G, Zhang D, Sun X, Hou B, Yu N (2023) TL-CNN-IDS: transfer learning-based intrusion detection system using convolutional neural network. J Supercomput 79(15):17562–17584. https://doi.org/10.1007/s11227-023-05347-4
9. Zhao R, Wang Y, Xue Z, Ohtsuki T, Adebisi B, Gui G (2023) Semisupervised federated-learning-based intrusion detection method for internet of things. IEEE Internet Things J 10(10):8645–8657. https://doi.org/10.1109/jiot.2022.3175918
10. Nandanwar H, Katarya R (2023) TL-BILSTM IOT: transfer learning model for prediction of intrusion detection system in IOT environment. Int J Inf Secur 23(2):1251–1277. https://doi.org/10.1007/s10207-023-00787-8
11. Hady AA, Ghubaish A, Salman T, Unal D, Jain R (2020) Intrusion detection system for healthcare systems using medical and network data: a comparison study. IEEE Access 8:106576–106584. https://doi.org/10.1109/access.2020.3000421
12. Clifton L, Clifton DA, Pimentel MA, Watkinson PJ, Tarassenko L (2014) Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. IEEE J Biomed Health Inform 18(3):722–730. https://doi.org/10.1109/jbhi.2013.2293059
13. Fotouhi H, Causevic A, Lundqvist K, Bjorkman M (2016) Communication and security in Health Monitoring Systems—a review. In: 2016 IEEE 40th annual computer software and applications conference (COMPSAC). https://doi.org/10.1109/compsac.2016.8
14. Rani AA, Baburaj E (2019) Secure and intelligent architecture for cloud-based healthcare applications in wireless body sensor networks. Int J Biomed Eng Technol 29(2):186. https://doi.org/10.1504/ijbet.2019.097305
15. Chakraborty, S., Aich, S., & Kim, H.-C. (2019). A secure healthcare system design framework using Blockchain technology. In: 2019 21st International Conference on Advanced Communication Technology (ICACT). https://doi.org/10.23919/icact.2019.8701983
16. Alabdulatif A, Khalil I, Forkan AR, Atiquzzaman M (2019) Real-time secure health surveillance for smarter health communities. IEEE Commun Mag 57(1):122–129. https://doi.org/10.1109/mcom.2017.1700547
17. Zhang J, Zulkernine M, Haque A (2008) Random-forests-based network intrusion detection systems. IEEE Trans Syst Man Cybern Part C Appl Rev 38(5):649–659. https://doi.org/10.1109/tsmcc.2008.923876
18. Rao BB, Swathi K (2017) Fast knn classifiers for network intrusion detection system. Indian J Sci Technol 10(14):1–10. https://doi.org/10.17485/ijst/2017/v10i14/93690
19. Shapoorifard H, Shamsinejad P (2017) Intrusion detection using a novel hybrid method incorporating an improved KNN. Int J Comput Appl 173(1):5–9. https://doi.org/10.5120/ijca2017914340
20. Tao H, Bhuiyan MZ, Abdalla AN, Hassan MM, Zain JM, Hayajneh T (2019) Secured data collection with hardware-based ciphers for IOT-based healthcare. IEEE Internet Things J 6(1):410–420. https://doi.org/10.1109/jiot.2018.2854714
21. Yaacoub J-PA, Noura M, Noura HN, Salman O, Yaacoub E, Couturier R, Chehab A (2020) Securing internet of medical things systems: Limitations, issues and recommendations. Futur Gener Comput Syst 105:581–606. https://doi.org/10.1016/j.future.2019.12.028
22. Saba T (2020) Intrusion detection in Smart City Hospitals using ensemble classifiers. In: 2020 13th International Conference on Developments in eSystems Engineering (DeSE). https://doi.org/10.1109/dese51703.2020.9450247
23. Thamilarasu G, Odesile A, Hoang A (2020) An intrusion detection system for internet of medical things. IEEE Access 8:181560–181576. https://doi.org/10.1109/access.2020.3026260
24. Kumar P, Gupta GP, Tripathi R (2021) An ensemble learning and fog-cloud architecture-driven cyber-attack detection framework for IOMT networks. Comput Commun 166:110–124. https://doi.org/10.1016/j.comcom.2020.12.003
25. Radoglou-Grammatikis P, Sarigiannidis P, Efstathopoulos G, Lagkas T, Fragulis G, Sarigiannidis A (2021) A self-learning approach for detecting intrusions in Healthcare Systems. IEEE Int Conf Commun. https://doi.org/10.1109/icc42927.2021.9500354

26. Zachos G, Essop I, Mantas G, Porfyrakis K, Ribeiro JC, Rodriguez J (2021) An anomaly-based intrusion detection system for internet of medical things networks. Electronics 10(21):2562. https://doi.org/10.3390/electronics10212562
27. Binbusayyis A, Alaskar H, Vaiyapuri T, Dinesh M (2022) An investigation and comparison of machine learning approaches for intrusion detection in IOMT Network. J Supercomput 78(15):17403–17422. https://doi.org/10.1007/s11227-022-04568-3
28. Saif S, Das P, Biswas S (2023) Dhoa-ANFIS: a hybrid technique to detect routing attacks in wireless body area network. Wirel Pers Commun 133(1):453–480. https://doi.org/10.1007/s11277-023-10774-w
29. Rathore H, Al-Ali AK, Mohamed A, Du X, Guizani M (2019) A novel deep learning strategy for classifying different attack patterns for deep brain implants. IEEE Access 7:24154–24164. https://doi.org/10.1109/access.2019.2899558
30. Manimurugan S, Al-Mutairi S, Aborokbah MM, Chilamkurti N, Ganesan S, Patan R (2020) Effective attack detection in internet of medical things smart environment using a deep belief neural network. IEEE Access 8:77396–77404. https://doi.org/10.1109/access.2020.2986013
31. Saheed YK, Arowolo MO (2021) Efficient cyber attack detection on the internet of medical things-smart environment based on deep recurrent neural network and machine learning algorithms. IEEE Access 9:161546–161554. https://doi.org/10.1109/access.2021.3128837
32. Awotunde JB, Abiodun KM, Adeniyi EA, Folorunso SO, Jimoh RG (2022) A deep learning-based intrusion detection technique for a secured IOMT system. Inform Intell Appl. https://doi.org/10.1007/978-3-030-95630-1_4
33. Khan S, Akhunzada A (2021) A hybrid DL-driven intelligent SDN-enabled malware detection framework for internet of medical things (IOMT). Comput Commun 170:209–216. https://doi.org/10.1016/j.comcom.2021.01.013
34. Nandy S, Adhikari M, Khan MA, Menon VG, Verma S (2022) An intrusion detection mechanism for secured IOMT framework based on Swarm-Neural Network. IEEE J Biomed Health Inform 26(5):1969–1976. https://doi.org/10.1109/jbhi.2021.3101686
35. Su J, Danilo Vasconcellos V, Prasad S, Daniele S, Feng Y, Sakurai K (2018) Lightweight classification of IOT malware based on image recognition. In: 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC). https://doi.org/10.1109/compsac.2018.10315
36. Nguyen H-T, Ngo Q-D, Le V-H (2018) IOT botnet detection approach based on psi graph and DGCNN classifier. In: 2018 IEEE International Conference on Information Communication and Signal Processing (ICICSP). https://doi.org/10.1109/icicsp.2018.8549713
37. Hussain F, Abbas SG, Fayyaz UU, Shah GA, Toqeer A, Ali A (2020) Towards a universal features set for IOT botnet attacks detection. In: 2020 IEEE 23rd International Multitopic Conference (INMIC). https://doi.org/10.1109/inmic50486.2020.9318106
38. Farhan RI, Maolood AT, Hassan NF (2020) Performance analysis of flow-based attacks detection on CSE-CIC-IDS2018 dataset using deep learning. Indonesian J Electr Eng Comput Sci 20(3):1413. https://doi.org/10.11591/ijeecs.v20.i3.pp1413-1418
39. Sarhan M, Layeghy S, Moustafa N, Portmann M (2021) NetFlow datasets for Machine Learning-based network intrusion detection systems. Lecture Notes Instit Comput Sci Soc Inform Telecommun Eng. https://doi.org/10.1007/978-3-030-72802-1_9
40. Saif S, Yasmin N, Biswas S (2023) Feature engineering based performance analysis of ML and DL algorithms for botnet attack detection in IOMT. Int J Syst Assur Eng Manag 14(S1):512–522. https://doi.org/10.1007/s13198-023-01883-7
41. Saif S, Das P, Biswas S, Khari M, Shanmuganathan V (2022) HIIDS: Hybrid intelligent intrusion detection system empowered with machine learning and metaheuristic algorithms for application in IOT based healthcare. Microprocess Microsyst. https://doi.org/10.1016/j.micpro.2022.104622
42. Reza S, Ferreira MC, Machado JJM, Tavares JM (2022) A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks. Expert Syst Appl 202:117275. https://doi.org/10.1016/j.eswa.2022.117275
43. Zhong M, Yi S, Fan J, Zhang Y, He G, Cao Y, Feng L, Tan Z, Mo W (2023) Power transformer fault diagnosis based on a self-strengthening offline pre-training model. Eng Appl Artif Intell 126:107142. https://doi.org/10.1016/j.engappai.2023.107142
44. Jiang H, Lin J, Kang H (2022) FGMD: A robust detector against adversarial attacks in the IOT network. Futur Gener Comput Syst 132:194–210. https://doi.org/10.1016/j.future.2022.02.019
45. Sachan S, Almaghrabi F, Yang J-B, Xu D-L (2021) Evidential reasoning for preprocessing uncertain categorical data for trustworthy decisions: an application on healthcare and finance. Expert Syst Appl 185:115597. https://doi.org/10.1016/j.eswa.2021.115597
46. Fonseca J, Bacao F (2023) Geometric smote for imbalanced datasets with nominal and continuous features. Expert Syst Appl 234:121053. https://doi.org/10.1016/j.eswa.2023.121053

47. Jain S, Pawar PM, Muthalagu R (2022) Hybrid intelligent intrusion detection system for internet of things. SSRN Electron J. https://doi.org/10.2139/ssrn.4097433
48. Fraihat S, Makhadmeh S, Awad M, Al-Betar MA, Al-Redhaei A (2023) Intrusion detection system for large-scale IOT NetFlow networks using machine learning with modified arithmetic optimization algorithm. Internet of Things 22:100819. https://doi.org/10.1016/j.iot.2023.100819
49. Sathish N, Varlarmathi K (2022) Detection of intrusion behavior in cloud applications using Pearson's chi-squared distribution and decision tree classifiers. Pattern Recogn Lett 162:15–21. https://doi.org/10.1016/j.patrec.2022.08.008
50. Nayak J, Meher SK, Souri A, Naik B (2022) Extreme learning machine and bayesian optimization-driven intelligent framework for IOMT cyber-attack detection. J Supercomput 78(13):14866–14891. https://doi.org/10.1007/s11227-022-04453-z
51. Mohy-eddine M, Guezzaz A, Benkirane S, Azrour M (2023) An intrusion detection model using election-based feature selection and K-NN. Microprocess Microsyst. https://doi.org/10.1016/j.micpro.2023.104966
52. Kumaran SS, Balakannan SP, Li J (2021) A deep analysis of object capabilities for intelligence considering wireless IOT devices with the DNN approach. J Supercomput 78(4):4745–4758. https://doi.org/10.1007/s11227-021-04064-0
53. Mishra S (2022) An optimized gradient boost decision tree using enhanced African buffalo optimization method for cyber security intrusion detection. Appl Sci 12(24):12591. https://doi.org/10.3390/app122412591
54. Mantas CJ, Castellano JG, Moral-García S, Abellán J (2018) A comparison of random forest based algorithms: random credal random forest versus oblique random forest. Soft Comput 23(21):10739–10754. https://doi.org/10.1007/s00500-018-3628-5
55. Tarbiyati H, Nemati Saray B (2023) Weight initialization algorithm for physics-informed neural networks using finite differences. Eng Comput. https://doi.org/10.1007/s00366-023-01883-y
56. Sai Chaitanya Kumar G, Kiran Kumar R, Parish Venkata Kumar K, Raghavendra Sai N, Brahmaiah M (2024) Deep residual convolutional neural network: an efficient technique for intrusion detection system. Expert Syst Appl 238:121912. https://doi.org/10.1016/j.eswa.2023.121912
57. Deo TY, Sanju A (2023) Data imputation and comparison of custom ensemble models with existing libraries like XGBoost, CATBoost, AdaBoost and Scikit learn for predictive equipment failure. Mater Today Proc 72:1596–1604. https://doi.org/10.1016/j.matpr.2022.09.410
58. Argus + ml. openargus. https://openargus.org/argus-ml
59. Dina AS, Siddique AB, Manivannan D (2023) A deep learning approach for intrusion detection in internet of things using focal loss function. Internet Things 22:100699. https://doi.org/10.1016/j.iot.2023.100699
60. Chaganti R, Mourade A, Ravi V, Vemprala N, Dua A, Bhushan B (2022) A particle swarm optimization and deep learning approach for intrusion detection system in internet of medical things. Sustainability 14(19):12828. https://doi.org/10.3390/su141912828
61. Wagan SA, Koo J, Siddiqui IF, Qureshi NM, Attique M, Shin DR (2023) A fuzzy-based duo-secure multi-modal framework for IOMT anomaly detection. J King Saud Univ Comput Inf Sci 35(1):131–144. https://doi.org/10.1016/j.jksuci.2022.11.007
62. Salman T, Ghubaish A, Unal D, Jain R (2020) Safety score as an evaluation metric for machine learning models of security applications. IEEE Netw Lett 2(4):207–211. https://doi.org/10.1109/lnet.2020.3016583
63. Gupta K, Sharma DK, Datta Gupta K, Kumar A (2022) A tree classifier based network intrusion detection model for internet of medical things. Comput Electr Eng 102:108158. https://doi.org/10.1016/j.compeleceng.2022.108158
64. Firat Kilincer I, Ertam F, Sengur A, Tan R-S, Rajendra Acharya U (2023) Automated detection of cybersecurity attacks in healthcare systems with recursive feature elimination and multilayer perceptron optimization. Biocybern Biomed Eng 43(1):30–41. https://doi.org/10.1016/j.bbe.2022.11.005
65. Dhanya L, Chitra R (2024) A novel Autoencoder based feature independent Ga optimised XGBoost classifier for IoMT malware detection. Expert Syst Appl 237:121618. https://doi.org/10.1016/j.eswa.2023.121618
66. Dhanya L, Chitra R (2023) An optimal differential evolution based XGB classifier for IoMT malware classification. In: 2023 International Conference on Advances in Intelligent Computing and Applications (AICAPS). https://doi.org/10.1109/aicaps57044.2023.10074030

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Prashant Giridhar Shambharkar is an Assistant Professor at the Department of Computer Science and Engineering at Delhi Technological University, Delhi, India. He has approximately 18 years of professional experience and has published more than 35 articles in reputed International Journals, Conferences, and Book chapters. Three Ph.D. scholars are currently pursuing Doctoral Degrees under his active supervision. His research interests include Machine Learning, Deep Learning, Blockchain, Natural Language Processing, and Data Mining. He is a reviewer of several reputed International Journals and Conferences.

Nikhil Sharma received B.Tech. degree from Guru Gobind Singh Indraprastha University, Delhi, India, in 2016, M.Tech. degree in Information Security from Ambedkar Institute of Advanced Communication Technologies and Research, Govt. of NCT of Delhi, India (Currently Netaji Subhash University of Technology East Campus, Delhi, India), in 2018. He is pursuing PhD in Computer Science and Engineering from Delhi Technological University, Delhi, India. His research interests include Machine Learning, Deep Learning, Blockchain, Internet of Things, and Information Security.

Optical and Quantum Electronics (2024) 56:1133
https://doi.org/10.1007/s11082-024-07076-1

Design and analysis of a rectangular core refractive index-based PCF sensor for bio-sensing application

Neeraj Singh[1] · Akash Khamaru[1] · Ajeet Kumar[1]

Received: 4 March 2024 / Accepted: 11 May 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
A hollow-core photonic crystal fiber based Refractive Index sensor in the terahertz frequency regime for bio-sensing application is proposed in this paper. Refractive index variation is used for the identification of different analytes with refractive indices ranging from n = 1.35 to 1.39. Zeonex is chosen as the fiber material. The proposed photonic crystal fiber (PCF) sensor features a solitary rectangular core enclosed by an arrangement of 16 rectangular air holes within the cladding area with different heights and widths. Simulation and numerical analysis are conducted using Finite Element Method-based based Comsol Multiphysics Software. The proposed PCF sensor exhibits relative sensitivity of 98.8%, an Effective Mode Area of 1.06×10^{-7} m^2 with a very low confinement loss of 7.751×10^{-14} dB/cm are achieved at a frequency of 1.4 THz for an analyte of refractive index 1.39 in X Polarization direction of electric field. Moreover, birefringence, non-linearity, effective refractive index and power fraction are also analyzed. Further simple PCF structure based on rectangle ensures that fabrication is feasible with current fabrication techniques.

Keywords Terahertz frequency regime · Rectangular core · Relative sensitivity · Biosensor

1 Introduction

Photonic crystal fibers are attracting attention in the terahertz (THz) range for their unique capabilities in guiding and manipulating electromagnetic waves. Moreover, high sensitivity, compact size, low cost, flexibility, and fast response make them ideal for a variety of sensing applications (Sultana et al. 2018). Particularly, bio-chemical sensing which includes detecting or sensing of various biological components such as different types of cancer cells, blood components, analysis of RNA, DNA, and proteins and malaria cells. Utilizing the low photon energy within the THz range, THz sensors offer exceptional penetration capability through porous samples while maintaining test layer integrity (Haddad et al. 2013). Furthermore, PCFs operating in the THz range offer

✉ Ajeet Kumar
ajeetdph@dtu.ac.in

[1] Advanced Photonics Simulation Research Laboratory, Department of Applied Physics, Delhi Technological University, Delhi, India

Published online: 02 June 2024

significant advantages in terms of sensing analyte due to their large filling area, as the diameter of air holes and fiber diameter is larger than PCFs that are being operated in the mid-infrared range (Ekhlasur Rahaman et al. 2020). Due to these properties, various THz sensors and detectors have been developed using PCF. The hexagonal, octagonal, elliptical, rectangular structure of photonic crystal fiber has already been introduced, with a hollow core or solid core. Hollow core fibers offer higher sensitivity than solid core fibers. Hollow-core fibers (HCFs) exhibit superior suitability for detecting distinct analytes having lower RI and also offer strong light-analyte interaction within the core (Hossain and Podder 2019). Compared to solid core or porous core PCF, HC-PCF has a larger core power fraction and a lower effective material loss since the core is only filled with the target analyte.

PCF is becoming an increasingly important component of contemporary biological and biomedical technology since it may be used to detect blood components and cancer cells in both invasive and non-invasive ways among other things. The traditional techniques, which are employed for various pathological examinations and imaging, take a very long time to perform and obtain findings. The terahertz radiation band has garnered significant attention from the scientific community in recent times as it is possible to transmit terahertz radiation using PCFs without any risk to human health (Ferdous et al. 2022). Numerous THz optical sensors have been documented for use in biological applications. The advantages of PCF enable researchers to test different models for the detection of different analytes. Arif et el. proposed a chemical sensor, having a hexagonal cladding with circular air holes for liquid sensing and attained a relative sensitivity value of 50%, 55.83%, and 59.07% for water, ethanol, and benzene, respectively (Arif et al. 2016). A PCF sensor based on terahertz regime suggested by Islam et al. achieved a relative sensitivity of 68.87% for Ethanol (Islam et al. 2018). A chemical sensor with rotated hexacore and heptagonal cladding design with circular holes proposed by Selim Hossain et al. exhibit a relative sensitivity of 68.48%, 69.20%, 66.78% for ethanol, benzene, and water analyte at 1 THz (Selim Hossain and Sen 2021). Rahaman et al. suggested a dumbbell shaped HC-PCF with rectangular air holes for chemical sensing using TOPAS as the fiber material and achieved a relative sensitivity of 96.25% for Ethanol (Rahaman et al. 2021). Hossain et al. investigated a rectangular core sensor for Bane chemical detection and achieved a relative sensitivity of 94.4% and a confinement loss of 1.71×10^{-14} cm^{-1} at frequency 1.8 THz (Hossain et al. 2020). Ahmed et al. proposed a blood component sensor for analytes such as RBCs, hemoglobin, WBCs, plasma, and water and achieved a sensitivity response of 80.93%, 80.56%, 80.13%, 79.91% and 79.39% at frequency 1.5 THz (Ahmed et al. 2019). Bulbul et al. investigated a general purpose biochemical sensor with rectangular air holes and obtained a relative sensitivity of 95.82% at 2.5 THz (Bulbul et al. 2021).

The proposed sensor is a refractive index sensor which means the difference in refractive index is used for the identification of different analytes with RI ranging from 1.35 to 1.39. Zeonex is used as the background material. Zeonex is a transparent thermoplastic. It is preferred due to its wide range of optical characteristics, including its consistent refractive index of 1.53 in the THz regime, low absorption loss of 0.2 cm^{-1}, high glass transition temperature, minimal water absorption, high biocompatibility, and exceptional chemical resistance even at high temperatures (Yadav et al. 2023). Rectangular air holes are introduced in the fiber with a rectangular hollow core in which analytes are filled. A full vector FEM is used for numerical analysis and COMSOL multiphysics software is used for simulation.

2 Sensor model design

Figure 1 displays the cross-sectional representation of the proposed PCF sensor. Zeonex is used as background materials with rectangular air holes and a rectangular core. The rectangular air holes of different widths and heights are arranged across PCF. The overall diameter of the proposed diameter is 3.3 mm. The Perfectly matched layer (PML) occupies the cladding area and serves as an artificial absorbing medium to absorb radiation energy at the outer surface in the computational area and the width of the PML is 165 µm. Within the domain, the complete mesh consists of 24,554 elements while along the boundary it comprises 2452 elements. The width W1 and height H1 of the core is 400 µm and 500 µm respectively. Width W2 is 200 µm and H2 is 1540 µm respectively. Parameters W3 and W4 are 1720 µm and 1520 µm respectively. Parameters H2, H3, H4 are 1540 µm, 1340 µm, 1240 µm respectively. The pitch (separation between two adjacent rectangles) is 20 µm. The core is infiltrated with analyte samples with refractive index ranges from 1.35 to 1.39 serving as a region where light analyte interaction takes place. To begin the study, firstly the core of the PCF sensor is injected with the liquid analyte. Figure 2 shows the light propagation through the core filled with analyte**.** Since very little light leaves the core region, there is very little loss. The red region exhibits the highest degree of light confinement, and as it approaches the core boundary, the light's intensity progressively drops which shows light well confined in core region. Furthermore, the direction of electric and magnetic field is shown by arrow. To measure the efficacy of the proposed sensor, vivid sensing properties such as relative senitivity, birefringence, effective mode area, and confinement loss are used as measures.

3 Fabrication feasibility

Various techniques like stack and draw (Bao et al. 2010), drilling, extrusion (Ghazanfari et al. 2017), sol–gel (Bise and Trevor 2005), capillary stacking, and 3D printing (Ebendorff-Heidepriem et al. 2014) are available for the fabrication of Photonic Crystal Fibers. The capillary stacking, stack and draw, and sol–gel fabrication techniques are ideal for circular-shaped air hole fabrication. Alternatively, asymmetrical PCF designs with

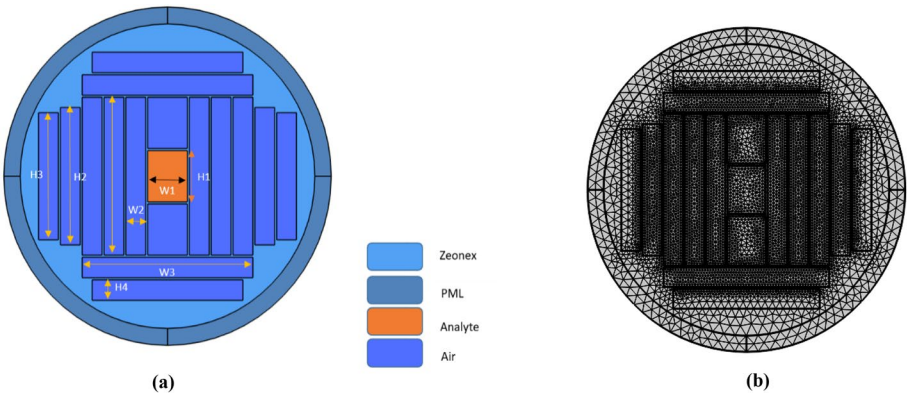


Fig. 1 **a** and **b** Sensor model design and meshing of proposed sensor

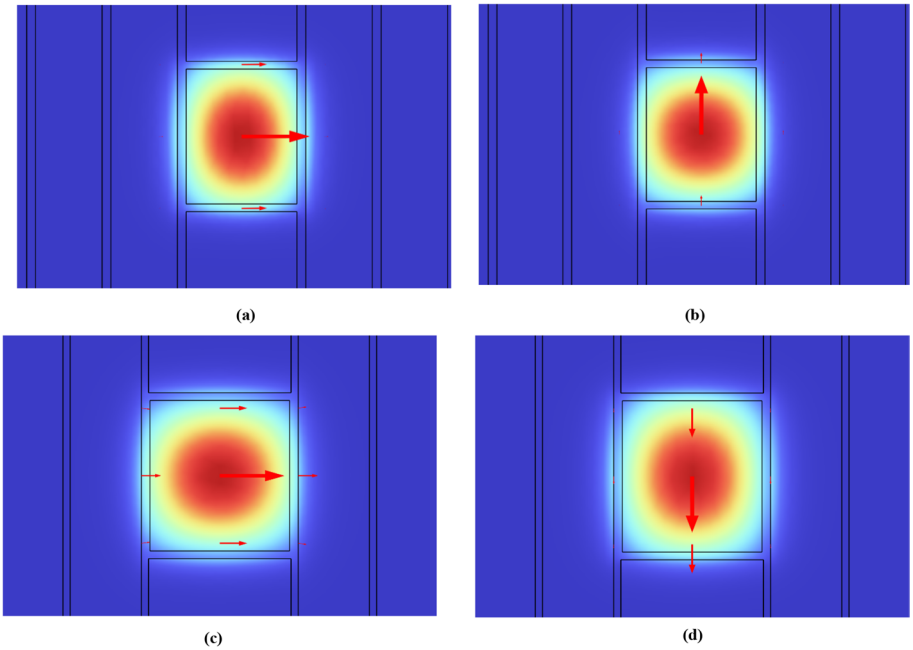


Fig. 2 **a**, **b** Mode electric field distribution and **c**, **d** Mode magnetic field distribution of the proposed fiber for analyte with refractive index 1.39 at 1.4 THz frequency in X and Y polarization direction respectively

rectangular, square, and elliptical air holes can be fabricated using 3D printing and extrusion technologies. Several asymmetrical PCF models have been fabricated by the Max Planck Institute, including elliptical and rectangular designs, demonstrating the successful fabrication of various PCF models (Bulbul et al. 2021). This ensures the feasibility of the fabrication of the suggested PCF fiber.

As the core of the proposed sensor is infiltrated with the analyte. Several infiltration techniques are available for the selective filing of PCF, which include selective filing by collapsing air holes (Xiao et al. 2004), by femtosecond laser micromachining (Wang et al. 2010), by splicing single-mode fiber etc.

4 Numerical analysis

The present study is done on three conditions Optimum (OPT), Optimum +2% (OPT +2%) and Optimum-2% (OPT −2%). To attain these conditions core width and height are altered by 2%. The width and height of the core are 400 µm and 500 µm for OPT condition. For OPT +2% condition the width and height of the core is 408 µm and 510 µm. Similarly, width and height have been set at 392 µm and 490 µm for condition OPT −2%. Different parameters are evaluated to analyze the performance of the proposed sensor for OPT, OPT +2%, OPT −2% conditions. These parameters include Relative sensitivity (RS), confinement loss (*CL*), power fraction (*P*), Non-linearity (γ), Effective mode area (EMA), and Birefringence. These parameters significantly impact the performance of the PCF sensors.

The RS is a crucial property that expresses the change in the optical response of the structure compared to the change in the external parameter. It measures the changes in the RI with the variation in the surroundings. It can be calculated using the following relation (Bulbul et al. 2022):

$$RS = \frac{n}{n_{eff}} \times P \tag{1}$$

Here, n is the RI of the sensing analyte and n_{eff} is the effective mode index (EMI). Here P represents the power fraction.

Power fraction (P) is the fraction of power located in the core of the fiber. It can be using the following equation (Habib et al. 2019),

$$P = \frac{sample \int R(E_x H_{y-} E_y H_x) dxdy}{whole \int R(E_x H_y - E_y H_x) dxdy} \times 100\% \tag{2}$$

Light traveling through the fiber experiences an average refractive index depending upon the design of the PCF. This average RI is called the effective mode index. EMI can be evaluated by Panda and Pukhrambam (2021):

$$n_{eff} = \frac{\beta}{k_o} \tag{3}$$

ere, β is the normalised propagation constant, k_o represents wave number in vacuum.

Birefringence is also one of the important parameter for designing the PCF sensor as high birefringence in PCF can be used to design sensors that are sensitive to the variation in RI. It describes the difference in refractive index between two perpendicular polarization directions (Yadav et al. 2023). It can be calculated to using,

$$B = |n_{eff}^x - n_{eff}^y| \tag{4}$$

The optimal region for analyte sensing and the cross-sectional area that interacts with the light is the effective mode area. We can evaluate the effective mode area by the relation:

$$EMA = \frac{\left(\iint |E|^2 dxdy \right)^2}{\iint |E|^4 dxdy} \tag{5}$$

Here, |E| denotes the electric field distribution of the proposed fiber sensor.

Confinement loss is an important factor in analyzing the efficiency of an optical device, particularly in photonic crystal fibers. It shows the amount of light leaked within an optical fiber or a waveguide due to inadequate light confinement within the core. Minimum the value of CL better would be the transmission of light within the waveguide.

$$CL\left(\frac{dB}{cm}\right) = 8.868 \times k_o \times l_m\left(n_{eff}\right) \times 10^{-2} \tag{6}$$

Here, k_o is the propagation constant, and $l_m(n_{eff})$ is the imaginary part of the effective refractive index of the propagating fundamental mode.

Non-linearity refers to the modifications in refractive index due to change in light intensity. By altering the parameters, a highly sensitive sensor can be designed to detect even subtle variations in the surrounding medium.

$$\Upsilon = \frac{2\pi}{\lambda}\left(\frac{n_2}{EMA}\right) \tag{7}$$

where n_2 is non-linear refractive index and EMA is effective mode area.

5 Result and discussion

To assess the guiding characteristics of the suggested rectangular core PCF, simulation and analysis are conducted using the FEM based COMSOL Multiphysics Software. Results for OPT, OPT − 2%, and OPT + 2% are evaluated and compared.

Figure 3 a, b shows OPT condition relative sensitivity variation with frequency for different analytes with RI ranging from n = 1.35 to 1.39. Relative sensitivity for analytes with higher refractive index is greater than for analytes with lower refractive index. We obtained relative sensitivity values 96.889%, 97.217%, 97.515%, 97.787%, 98.035% for RI of the analyte n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at an operating frequency of 1.4 THz. Relative sensitivity values of 97.868%, 98.135%, 98.378%, 98.599%, 98.800% are achieved for analyte with RI n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in X polarized direction at an operating frequency of 1.4 THz. These values indicate that there is an exceptionally high relative sensitivity obtained to all analyte samples.

Figure 4a, b shows relative sensitivity variation for OPT − 2% for different analytes. We obtained relative sensitivity values 94.997%, 95.481%, 95.921%, 96.321%, and 96.686% for RI of the analyte n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at an operating frequency of 1.4 THz. Relative sensitivity values of 96.05%, 96.478%, 96.864%, 97.213%, 97.529% are achieved for analyte with RI, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in X polarized direction at an operating frequency of 1.4 THz. These values indicate that there is an exceptionally high relative sensitivity obtained to all analyte samples.

Figure 5a, b shows relative sensitivity variation for OPT + 2%. We obtained relative sensitivity values 98.312%, 98.528%, 98.726%, 98.906%, 99.07% for RI of the analyte n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at an operating

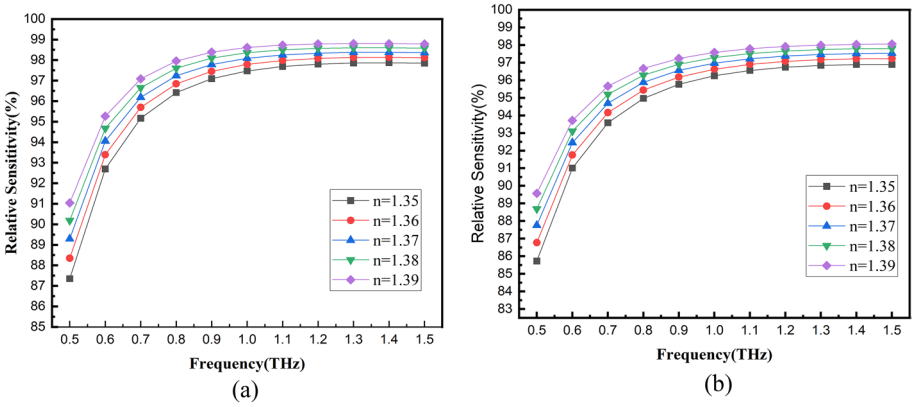


Fig. 3 **a** and **b** Relative sensitivity variation with frequency for OPT case in X and Y polarization directions respectively

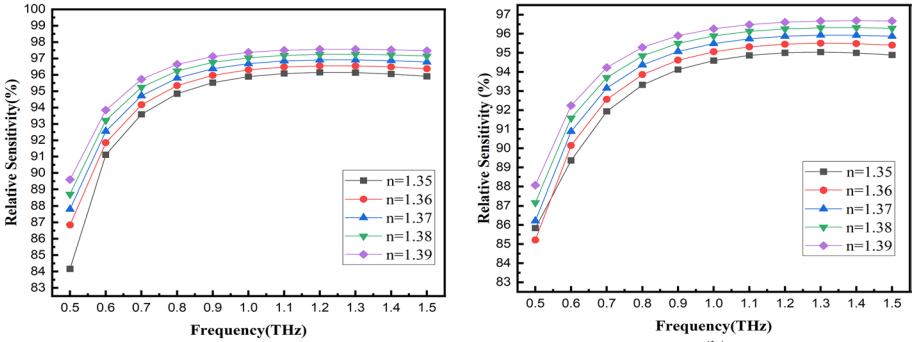


Fig. 4 **a** and **b** Relative sensitivity variation with frequency for OPT−2% case in X and Y polarization directions respectively

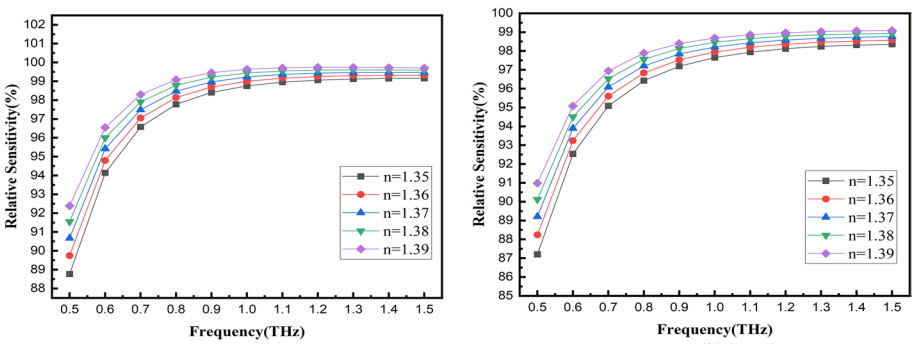


Fig. 5 **a** and **b** Relative sensitivity variation with frequency for OPT+2% case in X and Y polarization directions respectively

frequency of 1.4 THz. Relative sensitivity values of 99.149%, 99.315%, 99.466%, 99.604%, 99.730% are achieved for analyte with RI, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in X polarized direction at an operating frequency of 1.4 THz. These values indicate that there is an exceptionally high relative sensitivity obtained to all analyte samples.

Figure 6a–c shows birefringence of different analytes with RI, n = 1.35, 1.36, 1.37, 1.38, 1.39 for OPT, OPT +2%, OPT −2% condition. Birefringence decreases with increase in frequency. Birefringence values achieved for analytes with RI, n = 1.35, 1.36, 1.37, 1.38, 1.39 are 0.0015, 0.0015, 0.0014, 0.0014 and 0.0013 for OPT case; 0.0015, 0.0015, 0.0015, 0.0014, and 0.0013 for OPT −2% case; 0.0014, 0.0013, 0.0013, 0.0013, 0.0012 for OPT +2% case respectively at 1.4 THz operating frequency.

Figure 7a, b shows the EMA profile achieved for different analytes within the 1.35–1.39 RI range for X and Y polarization directions. With the increase in the frequency of light, there is a decrease in the EMA of the fiber. Higher frequencies mean shorter wavelengths and shorter wavelengths can be easily confined in the fiber. Furthermore, the fiber's waveguide dispersion becomes more prominent at higher frequencies, which leads to the tight confinement of light inside the fiber's core (Ferdous et al. 2022; Almawgani et al.

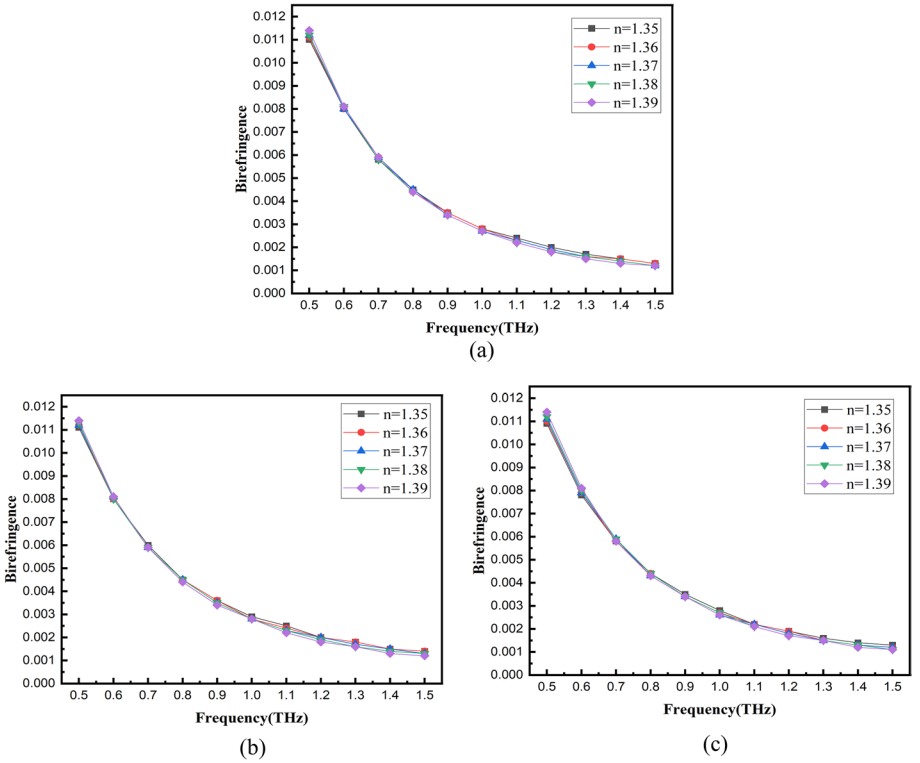


Fig.6 **a, b,** and **c** Birefringence variation with frequency for OPT, OPT−2%, OPT+2% respectively

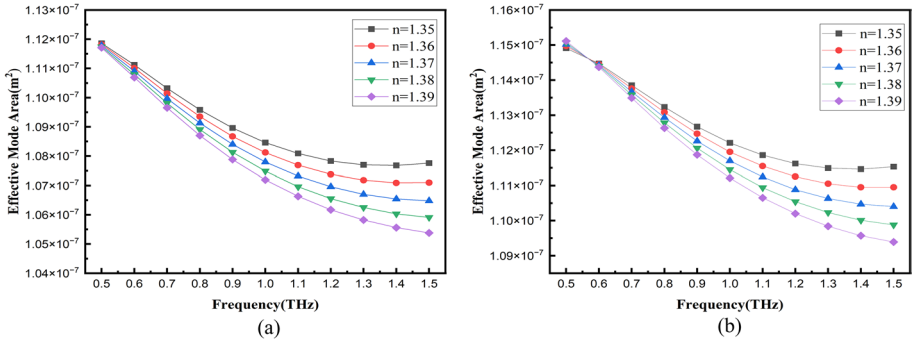


Fig. 7 **a** and **b** Effective mode area variation with frequency for OPT case in X and Y polarization directions respectively

2023). For OPT condition we achieved an average effective mode area of order 10^{-7} m^2. At 1.4 THz frequency effective mode area is 1.1146×10^{-7} m^2, 1.1095×10^{-7} m^2, 1.1047×10^{-7} m^2, 1.1001×10^{-7} m^2, 1.0957×10^{-7} m^2 for an analyte with RI, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively with Y polarization of electric field. In X polarized direction we observed an effective mode area of 1.0768×10^{-7} m^2, 1.0709×10^{-7} m^2, 1.0654×10^{-7} m^2,

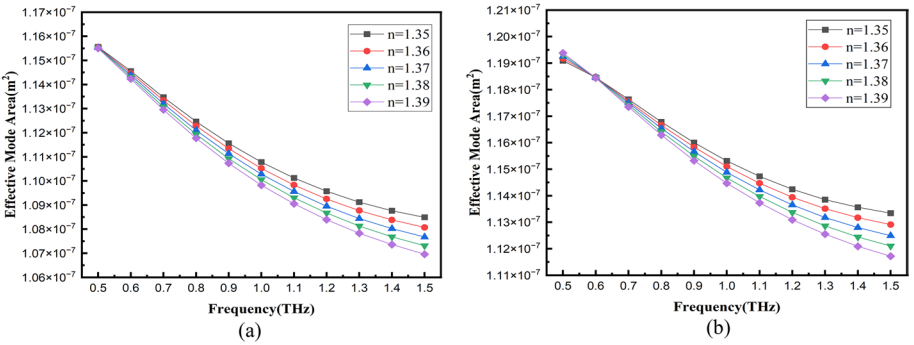


Fig. 8 **a** and **b** Effective mode area variation with frequency for OPT−2% case in X and Y polarization directions respectively

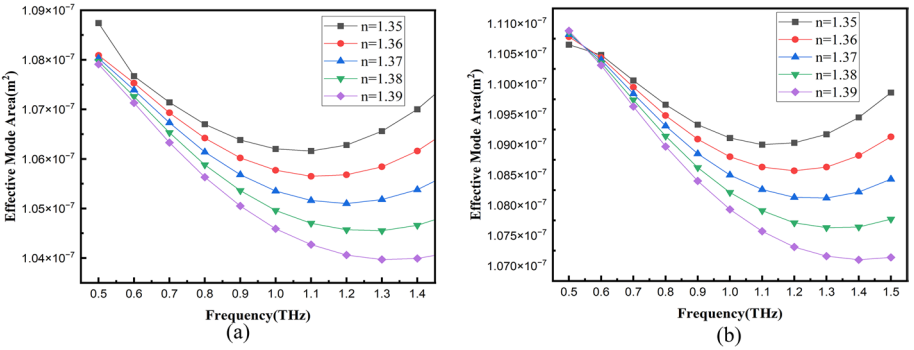


Fig. 9 **a** and **b** Effective mode area variation with frequency for OPT+2% case in X and Y polarization directions respectively

1.0603×10^{-7} m^2, 1.0556×10^{-7} m^2 for an analyte with RI, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively.

Figure 8a, b represents EMA for OPT −2% condition. For 1.4 THz frequency effective mode area is 1.0945×10^{-7} m^2, 1.0882×10^{-7} m^2, 1.0822×10^{-7} m^2, 1.0764×10^{-7} m^2, 1.0710×10^{-7} m^2 for an analyte with RI, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively with Y polarization of electric field. In X polarized direction we observed an effective mode area of 1.0700×10^{-7} m^2, 1.0616×10^{-7} m^2, 1.0538×10^{-7} m^2, 1.0466×10^{-7} m^2, 1.0399×10^{-7} m^2 for an analyte with RI, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively.

Figure 9a, b shows variation of EMAfor OPT +2% condition for different analytes. At 1.4 THz frequency effective mode area is 1.1356×10^{-7} m^2, 1.1317×10^{-7} m^2, 1.1280×10^{-7} m^2, 1.1244×10^{-7} m^2, 1.204×10^{-7} m^2 for an analyte with RI, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively with Y polarization of electric field. In X polarized direction we observed an effective mode area of 1.0876×10^{-7} m^2, 1.0838×10^{-7} m^2, 1.0802×10^{-7} m^2, 1.0768×10^{-7} m^2, 1.0736×10^{-7} m^2 for an analyte with RI, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively.

Figure 10a, b shows the non-linearity variation for OPT condition for both X and Y polarization directions in frequency range of 0.5–1.5 THz. Non-linearity increases linearly with frequency. We achieved non-linearity values of 5.528×10^{-9} W^{-1} m^{-1},

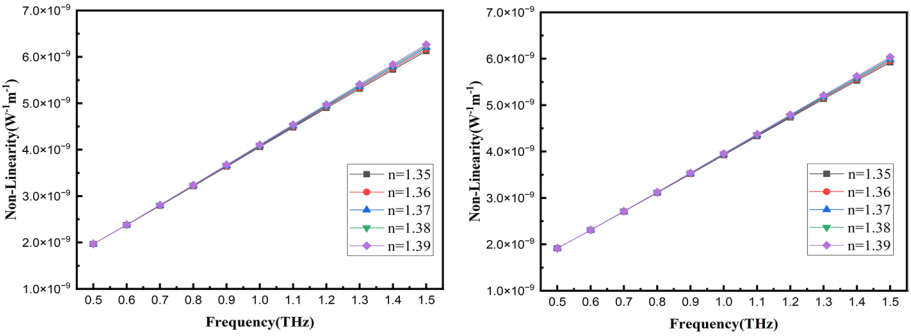


Fig. 10 **a** and **b** Non-Linearity variation with frequency for OPT case in X and Y polarization directions respectively

5.553×10^{-9} W^{-1} m^{-1}, 5.557×10^{-9} W^{-1} m^{-1}, 5.601×10^{-9} W^{-1} m^{-1}, 5.623×10^{-9} W^{-1} m^{-1} for RI of sensing analyte, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at 1.4 THz frequency. Non-linearity values of 5.72×10^{-9} W^{-1} m^{-1}, 5.753×10^{-9} W^{-1} m^{-1}, 5.783×10^{-9} W^{-1} m^{-1}, 5.811×10^{-9} W^{-1} m^{-1}, 5.837×10^{-9} W^{-1} m^{-1} were achieved for RI of sensing analyte n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in X polarized direction at 1.4 THz operating frequency.

As shown in Fig. 11a, b which is OPT −2% condition we achieved non-linearity values of 5.6297×10^{-9} W^{-1} m^{-1}, 5.6624×10^{-9} W^{-1} m^{-1}, 5.6939×10^{-9} W^{-1} m^{-1}, 5.67242×10^{-9} W^{-1} m^{-1}, 5.7533×10^{-9} W^{-1} m^{-1} for RI of sensing analyte, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at 1.4THz frequency. Non-linearity values of 5.7587×10^{-9} W^{-1} m^{-1}, 5.8041×10^{-9} W^{-1} m^{-1}, 5.8470×10^{-9} W^{-1} m^{-1}, 5.8875×10^{-9} W^{-1} m^{-1}, 5.9255×10^{-9} W^{-1} m^{-1} were achieved for RI of sensing analyte n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in X polarized direction at 1.4 THz operating frequency.

In Fig. 12a, b, for OPT +2% condition non-linear coefficient values of 5.4262×10^{-9} W^{-1} m^{-1}, 5.4448×10^{-9} W^{-1} m^{-1}, 5.4628×10^{-9} W^{-1} m^{-1},

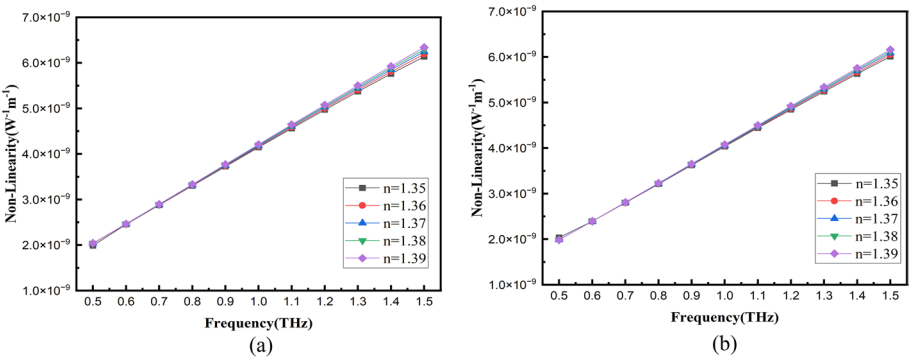


Fig. 11 **a** and **b** Non-Linearity variation with frequency for OPT−2% case in X and Y polarization directions respectively

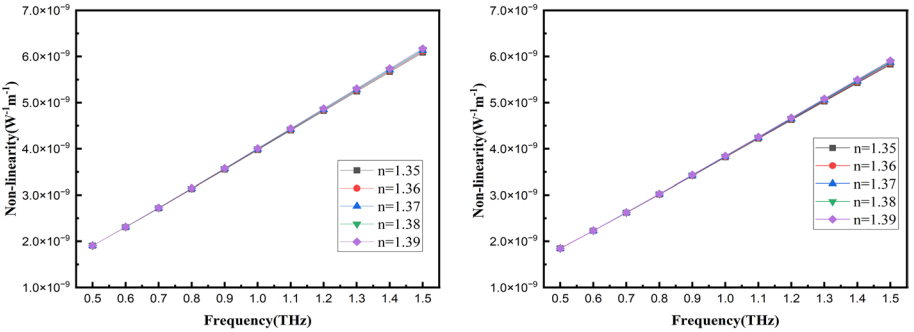


Fig. 12 **a** and **b** Non-Linearity variation with frequency for OPT +2% case in X and Y polarization directions respectively

5.4801×10^{-9} W^{-1} m^{-1}, 5.4970×10^{-9} W^{-1} m^{-1} for RI of sensing analyte, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at 1.4 THz frequency are achieved. Non-linear coefficient values of 5.6653×10^{-9} W^{-1} m^{-1}, 5.6854×10^{-9} W^{-1} m^{-1}, 5.7043×10^{-9} W^{-1} m^{-1}, 5.7223×10^{-9} W^{-1} m^{-1}, 5.7394×10^{-9} W^{-1} m^{-1} were achieved for RI of sensing analyte n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in X polarized direction at 1.4 THz operating frequency.

Figure 13a, b shows OPT condition CL variation for both X and Y polarization directions in the 0.5–1.5 THz frequency range. The CL graph depicts that confinement loss decreases as frequency increases. The confinement of lights gets better with larger frequencies as larger frequencies correspond to shorter wavelengths and shorter wavelengths are easier to confine within the fiber's periodic structure, the CL decreases for larger frequencies. Further with increase in frequency, the waveguide dispersion within the fiber becomes increasingly significant, which results in a more pronounced confinement of light within the core (Yadav et al. 2023). We observed the CL of 8.409×10^{-15} dB/cm, 8.659×10^{-15} dB/cm, 3.559×10^{-14} dB/cm, 3.849×10^{-14} dB/cm, 5.639×10^{-14} dB/cm for RI of analyte, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at 1.4 THz frequency. In X polarized direction we observed CL

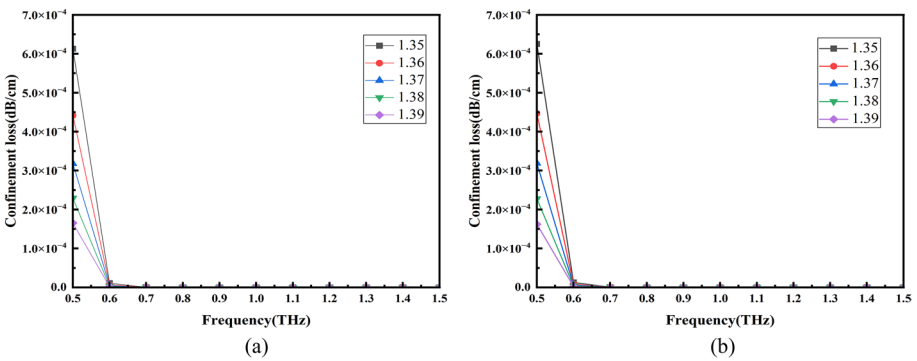


Fig. 13 **a** and **b** Confinement Loss variation with frequency for OPT case in X and Y polarization directions respectively

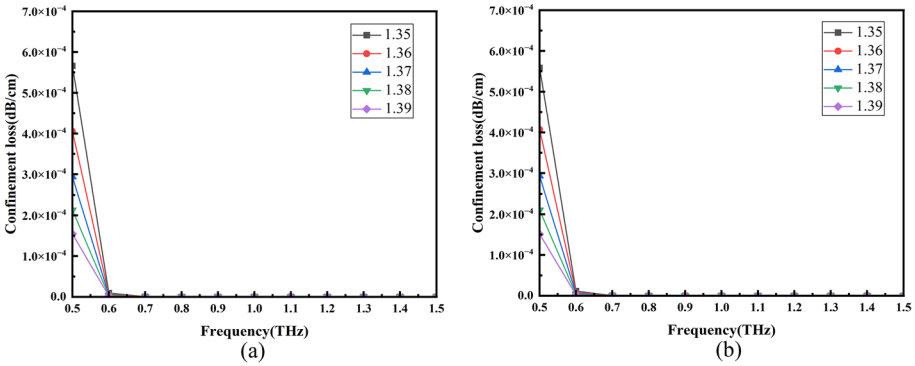


Fig. 14 **a** and **b** Confinement Loss variation with frequency for OPT −2% case in X and Y polarization directions respectively

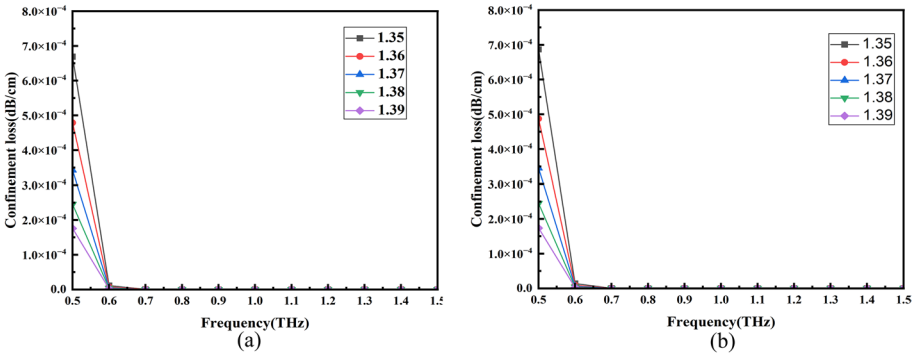


Fig. 15 **a** and **b** Confinement Loss variation with frequency for OPT−2% case in X and Y polarization directions respectively

of 2.779×10^{-14} dB/cm, 2.489×10^{-16} dB/cm, 1.499×10^{-13} dB/cm, 7.479×10^{-14} dB/cm, 1.759×10^{-14} dB/cm for the RI of analyte, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively at 1.4 THz frequency.

Figure 14a, b shows variation of CL for OPT −2% condition for which we achieved CL of 7.18×10^{-15} dB/cm, 1.51×10^{-14} dB/cm, 5.51×10^{-14} dB/cm, 9 .00× 10^{-14} dB/cm, 2.54×10^{-14} dB/cm for RI of analyte, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at 1.4 THz frequency. In X polarized direction we observed CL of 1.49×10^{-14} dB/cm, 9.56×10^{-13} dB/cm, 6.26×10^{-14} dB/cm, 6.55×10^{-14} dB/cm, 3.67×10^{-14} dB/cm for RI of analyte, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively at 1.4 THz frequency.

Figure 15a, b represents confinement loss profile for OPT +2% condition for which we achieved confinement loss of 6.53×10^{-14} dB/cm, 7.58×10^{-14} dB/cm, 9.62×10^{-14} dB/cm, 9.74×10^{-15} dB/cm, 1.48×10^{-13} dB/cm for refractive index of analyte, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at 1.4 THz frequency. In X polarized direction we observed confinement loss of 4.71×10^{-14} dB/cm, 1.68×10^{-14} dB/cm, 8.51×10^{-14} dB/cm, 5.05×10^{-14} dB/cm, 2.97×10^{-15} dB/cm for refractive index of analyte, n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively at 1.4 THz frequency.

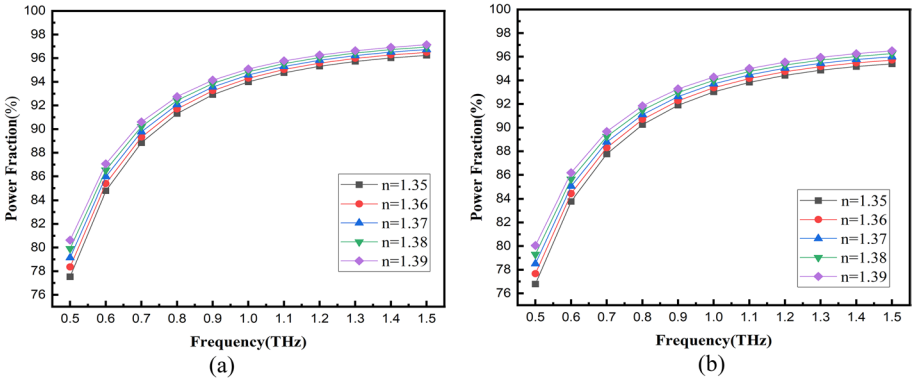


Fig. 16 **a** and **b** Power fraction variation with frequency for OPT case in X and Y polarization directions respectively

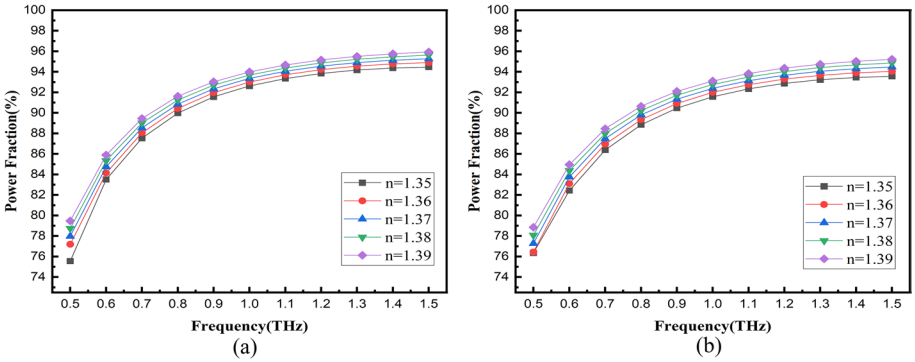


Fig. 17 **a** and **b** Power Fraction variation with frequency for OPT−2% case in X and Y polarization directions respectively

Figure 16a, b shows variation of power fraction of different analytes in X and Y polarized directions for OPT condition. One of the key factors influencing the sensor's performance is its sensitivity which is determined by Power fraction. The graph shows that core power fraction is generally more for analyte with higher refractive index. The values of power fraction have been achieved are 95.169%, 95.479%, 95.762%, 96.021%, 95.259% for RI of analyte n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at 1.4 THz frequency. We also observed power fraction values as 96.023%, 96.277%, 96.509%, 96.722%, 96.917% for RI of analyte n = 1.35, 1.36, 137, 1.38, 1.39 respectively in X polarized direction at operating frequency of 1.4 THz.

Figure 17a, b represents power fraction values for the OPT −2% condition. Power fraction values have been achieved are 93.445%, 93.894%, 94.304%, 94.678%, 95.020% for RI of analyte n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at 1.4THz frequency. We also observed power fraction values as 94.368%, 94.766%, 95.128%, 95.455%, 95.753% for RI of analyte n = 1.35, 1.36, 137, 1.38, 1.39 respectively in X polarized direction at operating frequency of 1.4 THz.

Figure 18a, b shows power fraction values for OPT +2% and we achieved power fraction values of 96.457%, 96.668%, 96.862%, 97.040%, 97.204% for RI of analyte n = 1.35, 1.36, 1.37, 1.38, 1.39 respectively in Y polarized direction at 1.4 THz frequency. We also observed power fraction values as 97.176%, 97.341%, 97.493%, 96.633%, 97.763% for RI of analyte n = 1.35, 1.36, 137, 1.38, 1.39 respectively in X polarized direction at operating frequency of 1.4 THz.

Table 1 shows the comparison between model suggested in this paper with previously suggested models in THz regime. The proposed design is a general-purpose sensor for bio sensing application which is capable of sensing different analytes with refractive index ranging from n = 1.35 to 1.39. Various analytes like Red blood cells, cancer cells have their refractive indices falling in RI range of 1.35–1.39, so the proposed detector is capable of detecting all these analytes. We have achieved better results than previously cited work with a design that is more sensitive to the RI change than many previously proposed designs.

6 Conclusion

In this paper, a rectangular core PCF sensor has been proposed for sensing different analyte with refractive index ranging from 1.35–1.39. The cladding is composed of rectangular shaped air holes. Zeonex is used as fiber material. The proposed refractive index fiber sensor can be utilized for detection of many biological components such as blood components (RBC, WBC, Plasma etc.), various cancer cells etc. The proposed sensor shows a very high relative sensitivity of 98.8% and 98.035% for an analyte of refractive index 1.39 at 1.4 THz operating frequency in X and Y polarized direction respectively. Besides CL is noticed only 7.75×10^{-14} dB/cm and 5.627×10^{-14} dB/cm in X and Y polarized direction respectively for an analyte of refractive index 1.39. Furthermore, other optical properties such as effective refractive index, EMA, and Non-linearity have been achieved. Table 2 shows values of different parameters for optimization conditions OPT, OPT −2%, OPT +2%. It is possible to build the suggested PCF using the current fabrication techniques. The proposed model has demonstrated a preference for being satisfactory in detecting various analytes.

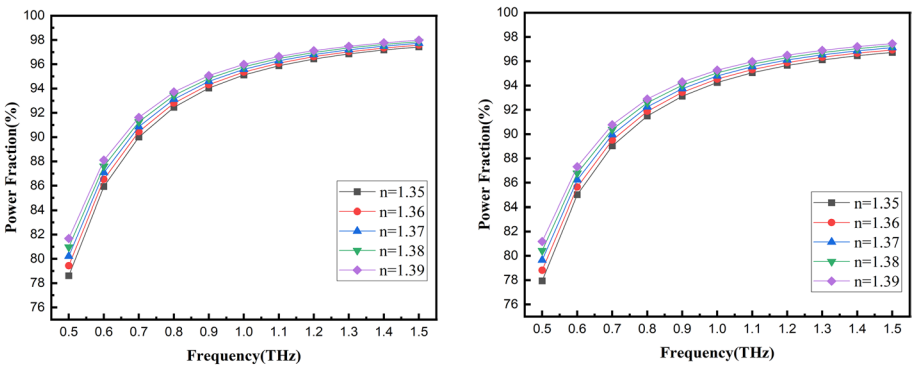


Fig. 18 **a** and **b** Power Fraction variation with frequency for OPT +2% case in X and Y polarization directions respectively

Table 1 Comparison of proposed fiber with existing photonic crystal fiber sensors

References	Detection target	Operating frequency (THz)	Background material	Effective mode area	Relative sensitivity (%)	Confinement loss
Hossain et al. (2020)	Blood components (RBCs)	1.5	Topas	1.55×10^{5} μm^2	80.93	–
Haddad et al. (2013)	Chemical (water)	1.6	Topas	1.57×10^{5} μm^2	_	1.06×10^{-9} cm^{-1}
Binti Suhaimi et al. (2020)	Chemical (liquid cocaine)	1	Zeonex	1.11×10^{5} μm^2	87.02	9.42×10^{-4} cm^{-1}
Mohammed et al. (2023)	Biosensor(tuberculosis)	2.2	Zeonex	4.342×10^{-8} m^2	90.60	3.13×10^{-9} cm^{-1}
This work	Biosensor	1.4	Zeonex	1.0559×10^{-7} m^2	98.80	1.759×10^{-14} dB/cm

Table 2 Different parameters for OPT, OPT −2% and OPT +2% cases at 1.4 THz for analyte with refractive index n = 1.39 in Y and X Polarization direction

Model	Analyte	Effective mode index		Effective mode area (m^2)		Non-linearity ($W^{-1}\ m^{-1}$)		Relative sensitivity (%)	
		Y	X	Y	X	Y	X	Y	X
OPT	n = 1.39	1.3648	1.3635	1.10×10^{-7}	1.06×10^{-7}	5.62×10^{-9}	5.84×10^{-9}	98.035	98.8
OPT −2%	n = 1.39	1.366	1.3647	1.07×10^{-7}	1.04×10^{-7}	5.75×10^{-9}	5.93×10^{-9}	96.686	97.529
OPT +2%	n = 1.39	1.3638	1.3626	1.12×10^{-7}	1.07×10^{-7}	5.50×10^{-9}	5.74×10^{-9}	99.07	99.73

The investigated PCF provides interesting findings and may find use in bio-sensing application. The proposed sensor can be easily fabricated with the help of the existing fabrication technology.

Author contributions N. S: Scientific computation, writing original draft, methodology, validation. A. Khumaru: Review and editing draft, validation, A. Kumar.: Supervision, conceptualization, visualization, editing draft

Funding The authors have not disclosed any funding.

Declarations

Competing interests The authors declare no competing interests.

References

Ahmed, K., Ahmed, F., Roy, S., et al.: Refractive index-based blood components sensing in terahertz spectrum. IEEE Sens. J. **19**, 3368–3375 (2019). https://doi.org/10.1109/JSEN.2019.2895166

Almawgani, A.H.M., Alhamss, D.N., Taya, S.A., et al.: Theoretical analysis of a refractive index sensor based on a photonic crystal fiber with a rectangular core. Opt. Quantum Electron. (2023). https://doi.org/10.1007/s11082-023-05172-2

Arif, M.F.H., Ahmed, K., Asaduzzaman, S., Azad, M.A.K.: Design and optimization of photonic crystal fiber for liquid sensing applications. Photonic Sens. **6**, 279–288 (2016). https://doi.org/10.1007/s13320-016-0323-y

Bao, H., Nielsen, K., Rasmussen. H. K. et al.: Fabrication and experimental characterization. (2010)

Binti Suhaimi, N.A.N., Yakasai, I.K., Abas, E., et al.: Modelling and simulation of novel liquid-infiltrated PCF biosensor in terahertz frequencies. IET Optoelectron. **14**, 411–416 (2020). https://doi.org/10.1049/iet-opt.2020.0069

Bise, R. T., Trevor, D. J.: Sol-gel derived microstructured fiber: fabrication and characterization. In: Optical fiber communication conference, vol. 3, p. 3. Optical Society of America, Anaheim (2005)

Bulbul, A.A.M., Kouzani, A.Z., Mahmud, M.A.P., Al, N.A.: Design and numerical analysis of a novel rectangular PCF (R-PCF)-based biochemical sensor (BCS) in the THz regime. Int. J. Opt. (2021). https://doi.org/10.1155/2021/5527724

Bulbul, A.A.M., Rahaman, H., Podder, E.: Design and quantitative analysis of low loss and extremely sensitive PCF-based biosensor for cancerous cell detection. Opt. Quantum Electron. (2022). https://doi.org/10.1007/s11082-022-03513-1

Ebendorff-Heidepriem, H., Schuppich, J., Dowler, A., et al.: 3D-printed extrusion dies: a versatile approach to optical material processing. Opt. Mater. Express **4**, 1494 (2014). https://doi.org/10.1364/ome.4.001494

Ekhlasur Rahaman, M., Bellal Hossain, M., Shekhar Mondal, H. et al.: Highly sensitive photonic crystal fiber liquid sensor in terahertz frequency range. In: Materials today: proceedings, pp. 3815–3820. Elsevier (2020)
Ferdous, A.H.M.I., Anower, M.S., Musha, A., et al.: A heptagonal PCF-based oil sensor to detect fuel adulteration using terahertz spectrum. Sens. Biosensing. Res. (2022). https://doi.org/10.1016/j.sbsr.2022.100485
Ghazanfari, A., Li, W., Leu, M.C., Hilmas, G.E.: A novel freeform extrusion fabrication process for producing solid ceramic components with uniform layered radiation drying. Addit. Manuf. **15**, 102–112 (2017). https://doi.org/10.1016/j.addma.2017.04.001
Habib, M.A., Anower, M.S., Abdulrazak, L.F., Reza, M.S.: Hollow core photonic crystal fiber for chemical identification in terahertz regime. Opt. Fiber Technol. (2019). https://doi.org/10.1016/j.yofte.2019.101933
El Haddad, J., Bousquet, B., Canioni, L., Mounaix, P.: Review in terahertz spectral analysis. TrAC—Trends Anal. Chem. **44**, 98–105 (2013)
Hossain, M.B., Podder, E.: Design and investigation of PCF-based blood components sensor in terahertz regime. Appl. Phys. A Mater. Sci. Process. (2019). https://doi.org/10.1007/s00339-019-3164-x
Hossain, M.B., Podder, E., Bulbul, A.A.M., Mondal, H.S.: Bane chemicals detection through photonic crystal fiber in THz regime. Opt. Fiber Technol. (2020). https://doi.org/10.1016/j.yofte.2019.102102
Islam, M.S., Sultana, J., Rifat, A.A., et al.: Terahertz sensing in a hollow core photonic crystal fiber. IEEE Sens. J. **18**, 4073–4080 (2018). https://doi.org/10.1109/JSEN.2018.2819165
Mohammed, N.A., Khedr, O.E., El-Rabaie, E.S.M., Khalaf, A.A.M.: High-sensitivity early detection biomedical sensor for tuberculosis with low losses in the terahertz regime based on photonic crystal fiber technology. Photonic Sens. (2023). https://doi.org/10.1007/s13320-023-0675-z
Panda, A., Pukhrambam, P.D.: Design and analysis of porous core photonic crystal fiber based ethylene glycol sensor operated at infrared wavelengths. J. Comput. Electron. **20**, 943–957 (2021). https://doi.org/10.1007/s10825-020-01650-y
Rahaman, M.E., Jibon, R.H., Mondal, H.S., et al.: Design and optimization of a PCF-based chemical sensor in THz regime. Sens. Biosens. Res. (2021). https://doi.org/10.1016/j.sbsr.2021.100422
Selim Hossain, M., Sen, S.: Design and performance improvement of optical chemical sensor based photonic crystal fiber (PCF) in the terahertz (THz) wave propagation. SILICON (2021). https://doi.org/10.1007/s12633-020-00696-8/Published
Sultana, J., Islam, Md.S., Ahmed, K., et al.: Terahertz detection of alcohol using a photonic crystal fiber sensor. Appl. Opt. **57**, 2426 (2018). https://doi.org/10.1364/ao.57.002426
Wang, Y., Liao, C.R., Wang, D.N.: Femtosecond laser-assisted selective infiltration of microstructured optical fibers. Opt. Express **18**(17), 18056–18060 (2010)
Xiao, L., Jin, W., Demokan, M.S., et al.: Fabrication of selective injection microstructured optical fibers with a conventional fusion splicer. Opt. Express **13**(22), 9014–9022 (2004)
Yadav, S., Lohia, P., Dwivedi, D.K.: A novel approach for identification of cancer cells using a photonic crystal fiber-based sensor in the terahertz regime. Plasmonics **18**, 1753–1769 (2023). https://doi.org/10.1007/s11468-023-01887-w

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3412712

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

Design of Dual-Band 4-Port Flexible MIMO Antenna for mm-Wave Technologies and Wearable Electronics

Rakesh N. Tiwari[1], Deepti Sharma[2], Prabhakar Singh[3] and Pradeep Kumar[4]

[1]Department of Electronics and Communication Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Andhara Pradesh-517325, India

[2]Department of Electronics and Communication Engineering, GL Bajaj Institute of Technology and Management, Greater Noida, Uttar Pradesh–201306 India

[3]Division of Physics, School of Basic Sciences, Galgotias University, Greater Noida, Uttar Pradesh 203201, India

[4]Discipline of Electrical, Electronic and Computer Engineering, University of KwaZulu-Natal, Durban 4041, South Africa

Corresponding author: Pradeep Kumar (pkumar_123@yahoo.com).

ABSTRACT In this paper, a dual-band 4-port Multiple Input Multiple Output (MIMO) antenna with very small footprints of 16×12×0.25 mm^3 is proposed for mm-wave applications. The proposed antenna operates in dual frequency bands at 28 GHz and 41.69 GHz. The isolation of the MIMO antenna is achieved >20dB at both the operating frequency bands by etching the slot at optimized position in the ground plane. The calculated diversity parameters of the antenna are acceptable for MIMO applications. The realized peak gain, Envelope Correlation Coefficient (ECC) and Diversity Gain (DG) are found to be 6.05 dBi, ≤0.10, and 9.89 dB, respectively. Moreover, the Mean Effective Gain (MEG), Channel Capacity Loss (CCL) and Total Active Reflection Coefficient (TARC) are evaluated to verify the proposed antenna performance in MIMO technology. The channel capacity is evaluated to demonstrate the superiority of the MIMO antenna over singe port antenna. Bending analysis of the antenna is also done to check the feasibility of this antenna for high data rate mm-wave wearable applications. In addition to that, antenna S-parameter is measured on chest and hand of human body for wireless body area network. Finally, to ensure the electromagnetic radiation safety, the specific absorption ratio (SAR) is evaluated at 28/41.69 GHz for input power of 500mW and found within the acceptable range.

INDEX TERMS Dual-band antenna, MIMO antenna, flexible antenna, diversity parameters, channel capacity, SAR, link margin, mm-wave and 5G applications

I. INTRODUCTION

Fifth generation communication systems are being implemented worldwide due to the demand of high data rate in diversified digital platforms [1, 2]. To fulfil the demand of the excessive data consumption by multiple users, millimeter frequency band is being utilized in the most of the countries. An adequate bandwidth at 28 GHz for 5G communication is of great interest for the mobile communication purpose. The specified 28 GHz band is mostly used due to relatively less absorption loss and path loss [3]. In addition to that, the multiple input multiple output (MIMO) antenna is one of the most promising designs for the transmission/receiving of quality signals [4]. In the modern wireless technology, MIMO antenna is widely used due to its spatial multiplexing, improved channel capacity, minimum multiple paths fading [5, 6]. Therefore, the effort has been made in [7] to design a dual-band MIMO antenna which consists of inset fed patch and parasitically coupled patch to operate at 28/38 GHz. In [8], the patch antenna with triangular stub was designed to obtain dual frequency response in millimeter range radiating at 28 and 38 GHz. Linearly arranged four patch antenna structure is reported covering the millimeter frequency bands 24 - 34 GHz and 37.5 - 41.5 GHz [9]. However, the overall linear dimension of the MIMO design is 48×12×0.254 mm^3. A crescent-shaped radiating patch is used to design a dual-band MIMO antenna for 5G application in mm-wave. The defected ground plane is used to improve the isolation, but the ground plane is not connected in this MIMO design [10]. Metamaterial based dual-band MIMO antennas are presented for better

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3412712

isolation in which decoupling structure is used between the two MIMO elements [11, 12]. In [13], 1×4 MIMO antenna is designed to operate at 38 GHz with self-isolated approach for 5G applications. A cross-shaped antenna is reported operating at 24.5 and 28.5 GHz [14]. This MIMO design consists of SSR as radiating element and demonstrates end-fire radiation. In [15], two antenna structures are designed on same ground plan to cover the sub 6 GHz band and mm-wave band simultaneously. A single element with multiple ports is also one of the approaches to design antenna for MIMO application as reported in [16]. This design operates at 28 GHz, but most of the diversity parameters of the MIMO antenna are not reported. The flexible MIMO antennas are another growing field of research in wireless industry and applicable in many contemporary fields of biotechnology, defense and sensors [17-19]. The various applications of the flexible antennas can be depicted in Fig. 1.

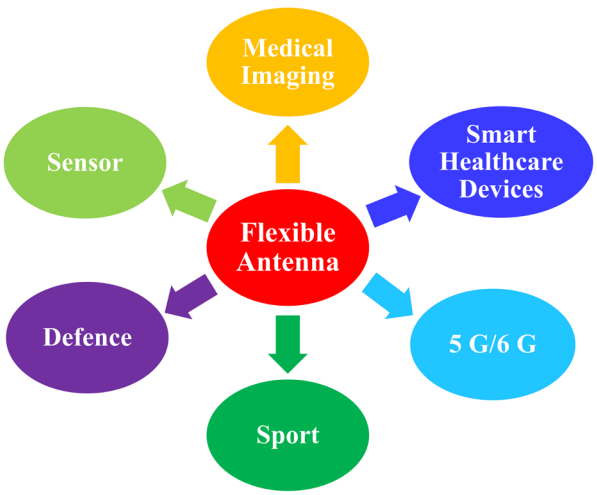


FIGURE 1. Flexible antenna applications.

In [20], two-port MIMO antenna is designed for sub-6 GHz ISM band applications. EBG structure is used to suppress the mutual coupling with overall gain of 9.50dBi. A 4×4 flexible MIMO antenna with size of 40×40×1.5 mm^3 is reported in [21]. The designed antenna demonstrates the dual-band response with UWB frequency range (from 3.18-12 GHz) by modifying the feedline. The antenna design for on-body application in the 28 GHz mm-wave band is reported in [22]. This design consists of a single radiating element with ground plane dimension of 29.06×11.42 mm^2. It is noticed that most of these flexible MIMO designs are operating in lower 5G bands.

Therefore, in this paper, a very small and flexible 4×4 MIMO antenna is reported for the millimeter frequency bands. The proposed antenna exhibits dual-band characteristics at 28/41.69 GHz. Various diversity parameters of the MIMO antenna are investigated using the CST Microwave Studio simulation software and measured results. In the existing literature, it is found that no one has reported bending analysis and analysed the channel capacity (CC) and SAR for the mm-wave 4-port MIMO antenna, which shows the significance of the proposed MIMO design.

The manuscript is organized in 10 sections including introduction as section 1. Sections 2 and 3 describe the antenna evolution and its implementation to design 4-port MIMO antenna. S-parameters and the important diversity parameters are evaluated and presented in sections 4 and 5, respectively. The CC and the bending analysis of the MIMO antenna are covered in sections 6 and 7, respectively. Moreover, to validate the use of proposed design for wearable electronics, the link margin is calculated in section 8, and on body measurement and SAR analysis are presented in section 9. Finally, the findings of the antenna design are concluded in section 10.

II. ANTENNA EVOLUTION AND S-PARAMETERS

The size miniaturization in the MIMO antenna is one of the important aspects for modern high speed communication systems. In this design, we have chosen the semi hexagonal microstrip patch (SHMP) because it occupies less area as compared to the semicircular patch, however, SHMP has current distribution similar to the semi disk patch.

The evolution stages of antenna design and the corresponding $|S_{11}|$ values are shown in Fig. 2 and Fig. 3, respectively. The SHMP was taken as the radiating patch along with a full ground plane for the initial antenna (Design-1). This antenna structure was fed using the microstrip line and resonates at 29.57 GHz. The resonance frequency of the design-1 is calculated by using certain approximations [23]:

(a) The SHMP is transformed to its equivalent rectangular patch of dimension $\mathcal{L} \times \mathcal{W}$.
(b) The rectangular patch is further transformed to its equivalent semicircular disk patch of effective radius ($\mathcal{A}_{eff}$).
(c) Now, using the value of $\mathcal{A}_{eff}$, the resonance frequency can be evaluated.

Thus, the design frequency of the proposed SHMP antenna can be given as:

$$\mathcal{F} = \frac{\xi_{mn} \times c}{2\pi \mathcal{A}_{eff} \sqrt{\varepsilon_{eff}}} \tag{1}$$

where,
$\xi_{11} = 1.84118$; for first dominant mode (TM_{11})
$\xi_{21} = 3.05424$; for second dominant mode (TM_{21})

$$\mathcal{A}_{eff} = \sqrt{\left\{2 \times \mathcal{L}_{eff} \times \left(\frac{\mathcal{W}_{eff}}{4}\right)\right\}} \tag{2}$$

Here, $\mathcal{L}_{eff}$, and $\mathcal{W}_{eff}$ are the effective length and width of the equivalent rectangular patch in which the fringing effect is taken into account [24]. ε_{eff} is the effective dielectric constant, and it can be calculated by taking the derived equivalent length and width of the rectangular patch as:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3412712

$$\mathcal{L} \approx 2 \times \mathcal{P} \tag{3}$$

$$\mathcal{W} \approx 3 \times \sqrt{3}\left(\frac{\mathcal{P}}{8}\right) \tag{4}$$

here,

$$\mathcal{P} \approx \sqrt{p_3 p_4}$$

Theoretically calculated values of lower and higher resonance frequencies are 29.44 and 48.84 GHz, respectively. When a narrow slot of dimension 2.62×0.31 mm^2 was etched in the ground plane (Design-2), the higher resonance is shifted towards lower side and the antenna radiates at two resonant frequencies of 25.55 and 41.69 GHz. Further, to obtain the lower resonance at 28 GHz, an additional stub was added on the top of SHMP (Design-3), while the higher resonance was stable at 41.69 GHz.

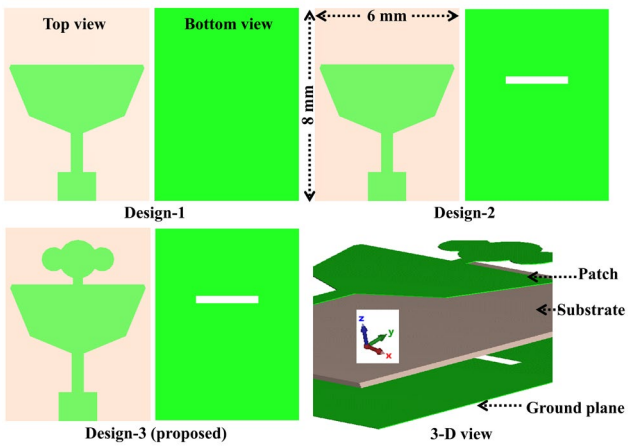


FIGURE 2. Antenna design stages.

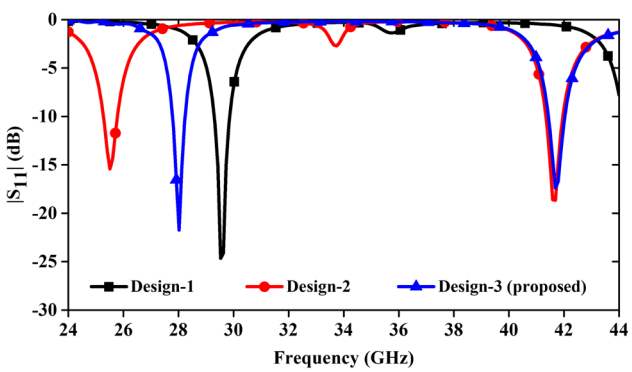


FIGURE 3. |S$_{11}$| graphs for different antenna designs.

III. 4-PORT MIMO ANTENNA DESIGN

The antenna design-3 is converted into a 4×4 MIMO antenna and demonstrated in Fig. 4. In this design, a flexible dielectric substrate (Rogers RO3003) with ε_r = 3.0, loss tangent 0.001, and thickness 0.25 mm is used. The overall dimension of the design is 16×12×0.25 mm^3. The optimized antenna dimensions are (in mm): L_s = 16, W_s = 12, p_1= 1.2, p_2 = 1.6, p_3 = 1.87, p_4 = 2.14, p_5 = 2.54, p_6 = 0.4, p_7 = 0.28, p_8 = 1.1, p_9 = 0.16, p_{10} = 0.5, p_{11} = 1.6, r_1 = 0.8, r_2 = 0.5, d = 0.52, r_2 = 1.2, L_g = 16, W_g = 12, g_1 = 2.62, g_2 = 0.31, g_3 = 5.6, g_4 = 3.38, g_5 = 4.89.

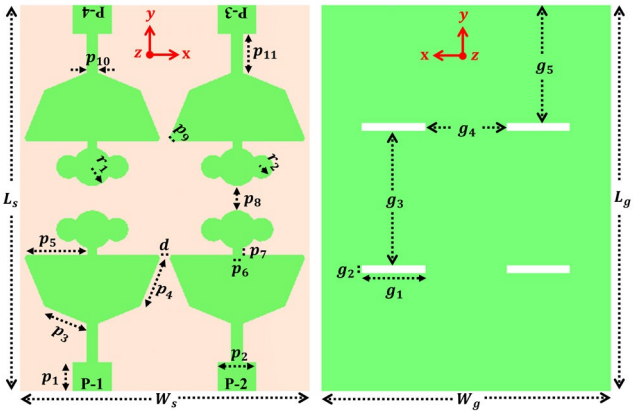


FIGURE 4. Structure of 4-port MIMO antenna.

IV. ANTENNA RESULTS AND DISCUSSION

The proposed MIMO antenna is fabricated and the prototype is shown in Fig. 5. The experimental setup for S-parameter measurement is depicted in Fig. 6. Vector Network Analyzer (VNA) with model no. N5247A is used to measure the scattering parameters of the MIMO antenna. The corresponding graph is shown in Fig. 7, in which the antenna shows the dual frequency bands at 27.6 and 40.0 GHz.

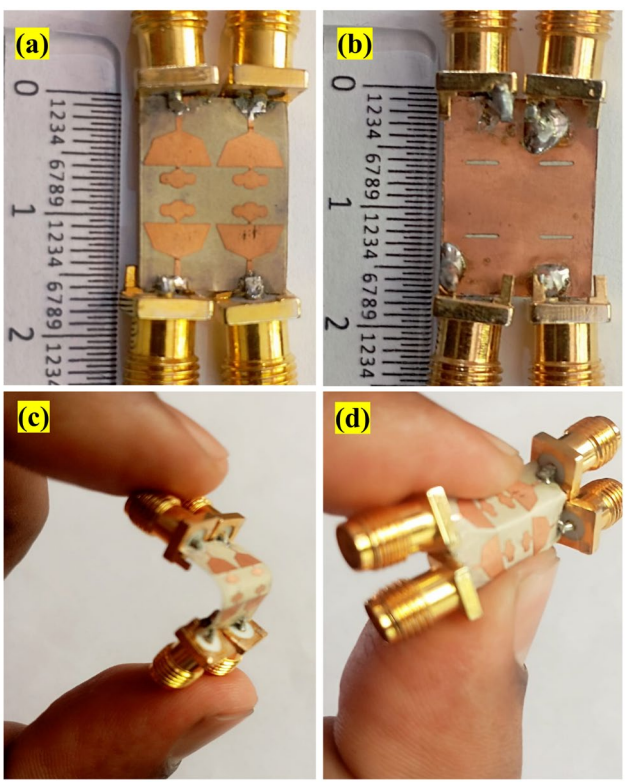


FIGURE 5. Prototype MIMO design, (a and b): flat configuration for top and bottom view, (c and d): deformation along x- and y- direction.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3412712

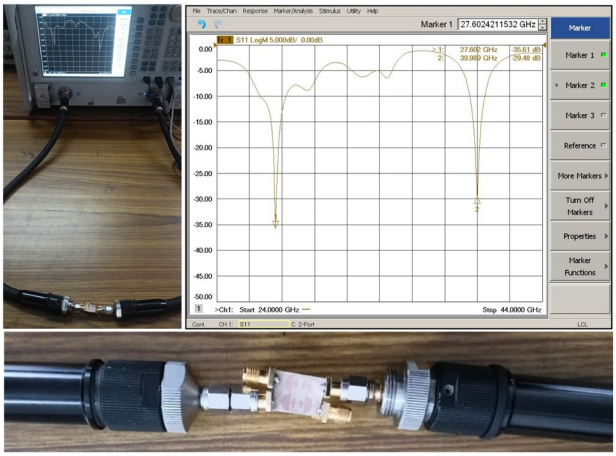


FIGURE 6. S-parameter measurement setup.

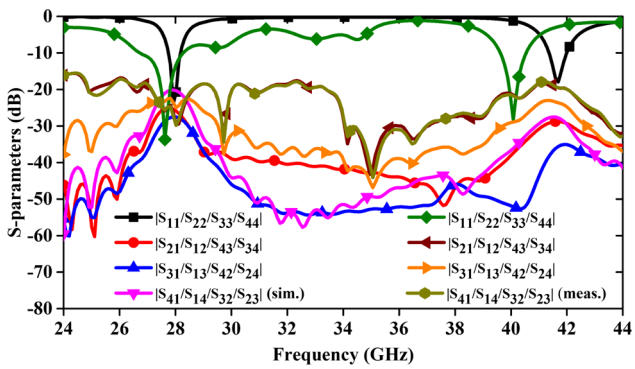


FIGURE 7. Scattering parameters of 4-port MIMO antenna.

The isolation across two frequency bands is >20dB, which is quite good for the MIMO applications. Some mismatch in the simulated and measured results may be due to imperfection in connector soldering and cable loss. Fig. 8 depicts the field density on the radiating elements. The current around the SHMP is maximum which indicates the fundamental lower resonance frequency. On the other hand, the upper resonance is developed due to higher mode. Thus, the resonance frequencies of the antenna can be easily tuned by varying the SHMP dimensions, stub loaded on the top of SHMP and the slots embedded in the ground plane.

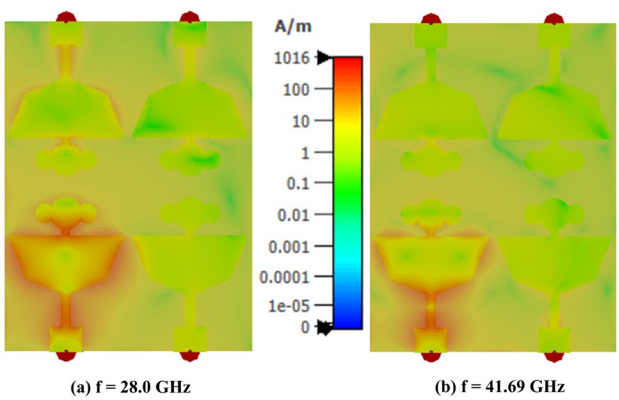


FIGURE 8. Current density variation at 28.0 and 41.69 GHz.

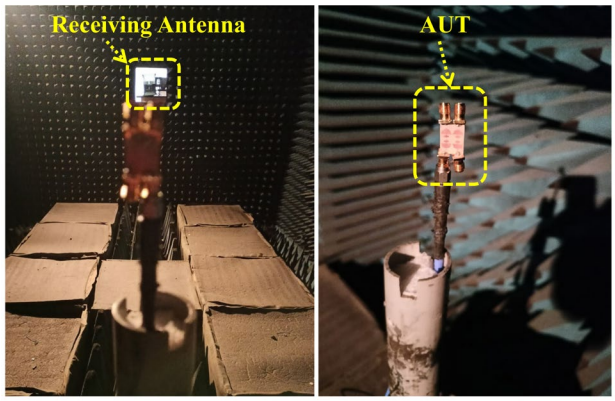


FIGURE 9. Gain measurement setup.

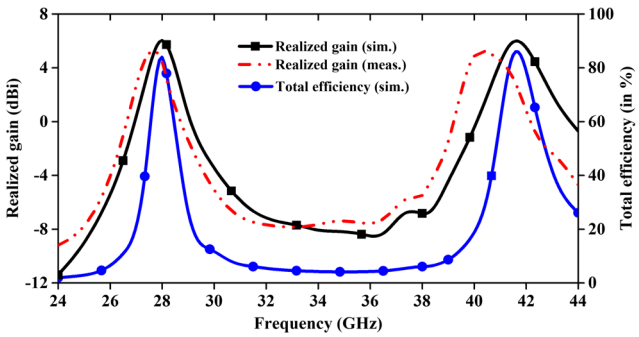


FIGURE 10. Realized gain and total efficiency.

The gain measurement setup is depicted in Fig. 9. The realized gain and total radiation efficiency (Fig. 10) at both the bands are above 6.0dBi and 80%, respectively. Furthermore, the co-and cross-polar E-and H-plane radiation patterns are presented in Fig. 11. The radiation patterns at 28 GHz are plotted for $\phi = 90^0$ and 0^0 and from the curve it is seen that the co-and cross-polarization discrimination is quite acceptable.

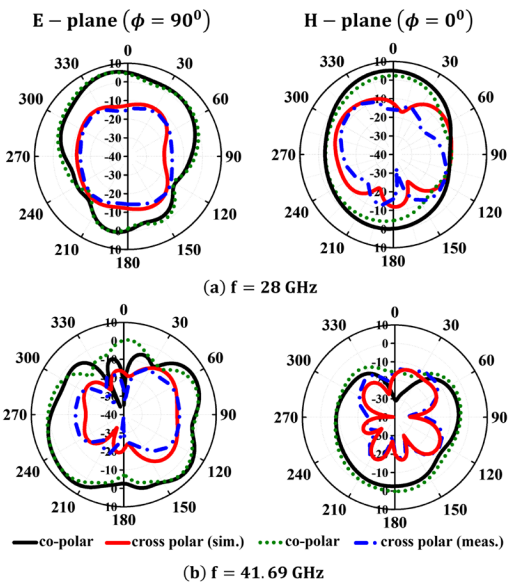


FIGURE 11. Radiation patterns of the proposed antenna.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3412712

However, at 41.69 GHz the polarization discrimination is > -15dB at the elevation angle 180^0. Some ripples in the radiation patterns may be due to the interference from the electromagnetic waves generated in coupled resonators of MIMO antenna.

V. DIVERSITY PARAMETERS

To access the quality performance of the MIMO antenna, the diversity parameters need to be investigated. In this section, the results of various diversity parameters are evaluated and discussed.

A. ENVELOPE CORRELATION COEFFICIENT (ECC) AND DIVERSITY GAIN (DG)

The ECC (Υ_{ecc}) and DG are evaluated across both the operating bands and shown in Fig. 12. The ECC value is calculated using 3-D radiation patterns [25]

$$\Upsilon_{ecc(pq)} = \frac{\left|\int_0^{2\pi}\int_0^{\pi}\{\mathcal{X}_{pr}\cdot\mathcal{E}_{\vartheta p}\mathcal{E}^*_{\vartheta q}\mathcal{P}_\vartheta + \mathcal{E}_{\varphi p}\mathcal{E}^*_{\varphi q}\mathcal{P}_\varphi\}d\Omega\right|^2}{\mathcal{M}\times\mathcal{N}} \quad (5)$$

here,

$$\mathcal{M} = \int_0^{2\pi}\int_0^{\pi}\{\mathcal{X}_{pr}\cdot\mathcal{E}_{\vartheta p}\mathcal{E}^*_{\vartheta p}\mathcal{P}_\vartheta + \mathcal{E}_{\varphi p}\mathcal{E}^*_{\varphi p}\mathcal{P}_\varphi\}d\Omega$$

$$\mathcal{N} = \int_0^{2\pi}\int_0^{\pi}\{\mathcal{X}_{pr}\cdot\mathcal{E}_{\vartheta q}\mathcal{E}^*_{\vartheta q}\mathcal{P}_\vartheta + \mathcal{E}_{\varphi q}\mathcal{E}^*_{\varphi q}\mathcal{P}_\varphi\}d\Omega$$

and,

$$\mathcal{X}_{pr} = \frac{\mathcal{P}_v}{\mathcal{P}_h} = 1$$

In ideal condition the totally uncorrelated ECC value of the MIMO antenna is zero and DG is 10 dB but the experimental value of ECC < 0.5 is satisfactory. Using these equations Υ_{ecc} found to be ≤ 0.20 at two resonance frequencies. Moreover, the DG is above 9.89 dB across two resonating bands.

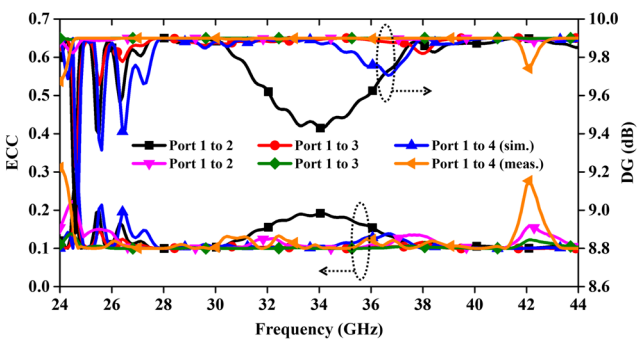


FIGURE 12. ECC and DG values of MIMO antenna for different ports.

B. MEAN EFFECTIVE GAIN (MEG)

MEG ($\psi_{\mathcal{M}}$) is the ratio of accepted mean power to the average incident power by the elements of MIMO antenna as compared to an isotropic antenna. The gain performance of the MIMO antenna can be analyzed considering the atmospheric effect and it is mathematically expressed as [26]

$$\psi_{\mathcal{M}} = \frac{\mathcal{P}_r}{\mathcal{P}_{in}} = \int_0^{2\pi}\int_0^{\pi}\left[\frac{\mathcal{X}_{pr}}{\mathcal{X}_{pr}+1}\times\mathcal{F}_\vartheta(\vartheta,\varphi)\mathcal{P}_\vartheta(\vartheta,\varphi) + \frac{1}{1+\mathcal{X}_{pr}} \times\mathcal{F}_\varphi(\vartheta,\varphi)\mathcal{P}_\varphi(\vartheta,\varphi)\right]\sin\vartheta d\vartheta d\varphi \quad (6)$$

where, $\mathcal{F}_\vartheta$ and $\mathcal{F}_\varphi$ are the power gain patterns of the MIMO antenna elements. The values of $\mathcal{F}_\vartheta$ and $\mathcal{F}_\varphi$ are calculated by varying the angle ϑ and keeping the angle φ constant and vice-versa. The MEG values are calculated and measured for the different ports (Fig. 13). It is to be noted that the difference of MEG between these ports must be <-3dB in 4-port MIMO antenna structure. From this figure, $\psi_{\mathcal{M}}$ values of all 4-ports are <-6dB across the dual frequency bands (simulated and measured).

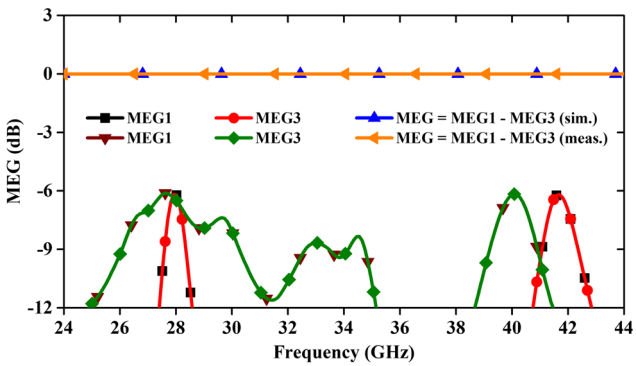


FIGURE 13. MEG against frequency.

C. CHANNE CAPACITY LOSS (CCL)

CCL ($\mathcal{C}_{\mathcal{CL}}$) is the parameter which measures the signal distortion between the transmitted and receiving end. The $\mathcal{C}_{\mathcal{CL}}$ is mathematically expressed as [27]

$$\mathcal{C}_{\mathcal{CL}} = -\log_2 det(\zeta) \quad (7)$$

where,

$$\zeta = \begin{bmatrix}\zeta_{11} & \zeta_{12} & \zeta_{13} & \zeta_{14}\\ \zeta_{21} & \zeta_{22} & \zeta_{23} & \zeta_{24}\\ \zeta_{31} & \zeta_{32} & \zeta_{33} & \zeta_{34}\\ \zeta_{41} & \zeta_{42} & \zeta_{43} & \zeta_{44}\end{bmatrix}$$

here,

$$\zeta_{pp} = 1 - \sum_q^4 |S_{pq}|^2$$

and,

$$\zeta_{pq} = -\left(S^*_{pp}S_{pq} + S^*_{qp}S_{pq}\right)$$

The simulated and measured CCL values are depicted in Fig. 14, and the respective values are 0.31 and 0.28 and, 0.25 and 0.22 bits/sec/Hz across both the operating bands. The simulated and measured value is always ≤ 0.4 bits/sec/Hz.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3412712

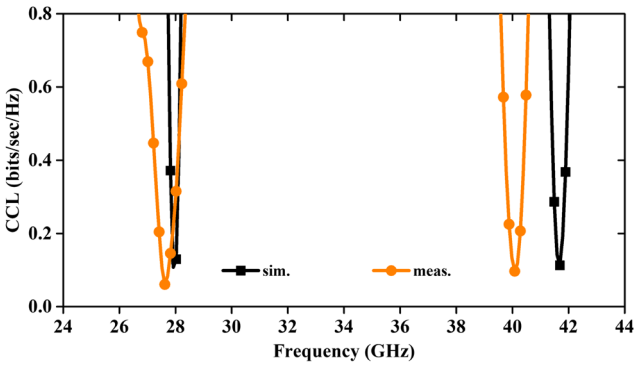


FIGURE 14. CCL against frequency.

D. TOTAL ACTIVE REFLECTION COEFFICIENT (TARC)

The TARC $(\mathcal{T}_\chi)$ is another parameter to express the operational bands of the MIMO antenna when the phase of the input signal is changed. The $\mathcal{T}_\chi$ of the 4-port MIMO antenna can be expressed as [28]

$$\mathcal{T}_\chi = \sqrt{\sum_{i=1}^{4}|\mathcal{R}_{in}|^2} \Big/ \sqrt{\sum_{i=1}^{4}|\mathcal{I}_{in}|^2} \tag{8}$$

where, $\mathcal{I}_{in}$ = incident signal, $\mathcal{R}_{in}$ = reflected signal.

The scattering matrix of the 4×4 mm-wave MIMO antenna can be expressed as

$$\begin{pmatrix}\mathcal{R}_1\\ \mathcal{R}_2\\ \mathcal{R}_3\\ \mathcal{R}_4\end{pmatrix} = \begin{pmatrix}\mathcal{S}_{11} & \mathcal{S}_{12} & \mathcal{S}_{13} & \mathcal{S}_{14}\\ \mathcal{S}_{21} & \mathcal{S}_{22} & \mathcal{S}_{23} & \mathcal{S}_{24}\\ \mathcal{S}_{31} & \mathcal{S}_{32} & \mathcal{S}_{33} & \mathcal{S}_{34}\\ \mathcal{S}_{41} & \mathcal{S}_{42} & \mathcal{S}_{43} & \mathcal{S}_{44}\end{pmatrix}\begin{pmatrix}\mathcal{I}_1\\ \mathcal{I}_2\\ \mathcal{I}_3\\ \mathcal{I}_4\end{pmatrix} \tag{9}$$

From equations (8) and (9), we can evaluate the value of $\mathcal{T}_\chi$ for the proposed antenna as

$$\mathcal{T}_\chi = \sqrt{\sum_{i=1}^{4}\left|\mathcal{S}_{i1} + \sum_{p=2}^{4}\mathcal{S}_{ip}e^{\jmath\vartheta_{p-1}}\right|^2} \Big/ \sqrt{4} \tag{10}$$

$$\mathcal{T}_\chi = \sqrt{\mathbb{A}+\mathbb{B}+\mathbb{C}+\mathbb{D}}/\sqrt{4} \tag{11}$$

here,

$$\mathbb{A} = \left|\mathcal{S}_{11} + S_{12}e^{\jmath\vartheta_1} + S_{13}e^{\jmath\vartheta_2} + S_{14}e^{\jmath\vartheta_3}\right|^2$$

$$\mathbb{B} = \left|\mathcal{S}_{21} + S_{22}e^{\jmath\vartheta_1} + S_{23}e^{\jmath\vartheta_2} + S_{24}e^{\jmath\vartheta_3}\right|^2$$

$$\mathbb{C} = \left|\mathcal{S}_{31} + S_{32}e^{\jmath\vartheta_1} + S_{33}e^{\jmath\vartheta_2} + S_{34}e^{\jmath\vartheta_3}\right|^2$$

$$\mathbb{D} = \left|\mathcal{S}_{41} + S_{42}e^{\jmath\vartheta_1} + S_{43}e^{\jmath\vartheta_2} + S_{44}e^{\jmath\vartheta_3}\right|^2$$

The measured $\mathcal{T}_\chi$ of the MIMO antenna is depicted in Fig. 15. Here, phase of the first port signal is fixed and the rest of the three ports are excited with different phase angles in between 0^0 to 180^0. It is found that $\mathcal{T}_\chi$ of the antenna is least deviated with the change of phase angles and hence the antenna elements are properly decoupled.

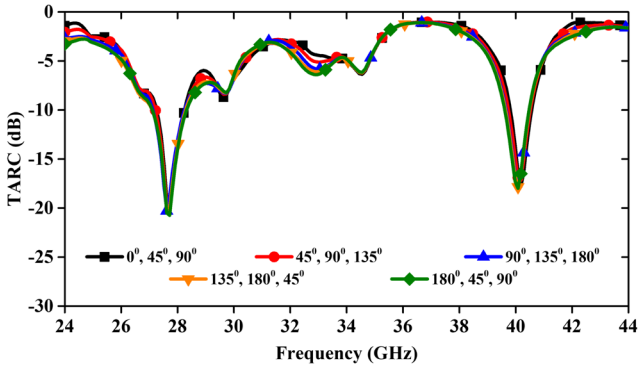


FIGURE 15. TARC against frequency at different signal phase angles.

VI. CHANNEL CAPACITY (CC)

The CC for this MIMO design is calculated using [29].

$$CC = log_2\left[det\left\{\mathcal{I}_{n_{\mathcal{R}_x}} + \frac{(\mathcal{SNR})_{avg.}}{n_{\mathcal{T}_x}}\mathcal{H}\mathcal{H}^\zeta\right\}\right] \tag{12}$$

here, $\mathcal{H}$ and $\mathcal{H}^\zeta$ indicate the normalized channel matrix, and its conjugate transpose. $\mathcal{I}_{n_{\mathcal{R}_x}}$ is the identity matrix and $n_{\mathcal{T}_x}$, $n_{\mathcal{R}_x}$ are the number of transmitted and received antenna elements.

The CC of the single input single output (SISO) and MIMO (in ideal case) are compared with the designed MIMO antenna in Fig. 16. Here CC is plotted at the signal to noise ratio (SNR) for 20 dB (Fig. 16). From the curve we can find that a SISO antenna, 4×4 MIMO structure and the proposed 4×4 MIMO antenna exhibit the CC of 5.85, 22.15, and 20.86 bps/Hz, respectively. Thus, it can be concluded that the CC of the reported 4×4 MIMO antenna is almost four times as compared to SISO antenna. The cumulative distribution function (CDF) is calculated at SNR value of 20 dB and it is shown in Fig. 17. From this figure, it can be observed that the designed mm-wave MIMO antenna is higher CC than SISO antenna. This antenna is more effective for high data rate 5G and beyond 6G technology applications.

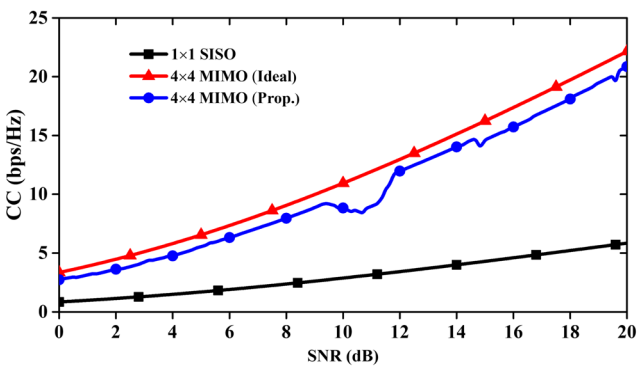


FIGURE 16. Channel capacity.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3412712

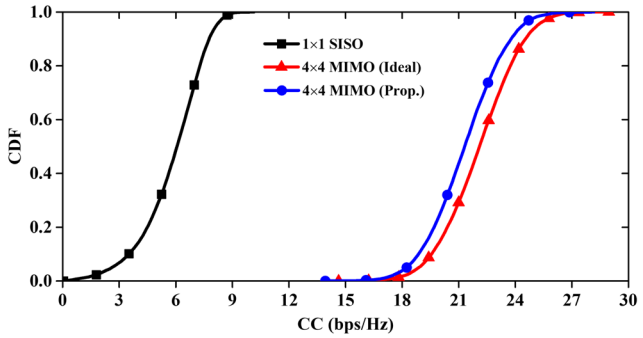


FIGURE 17. CDF against channel loss at SNR 20 dB.

VII. BENDING ANALYSIS OF MIMO ANTENNA

In this section, an analysis is presented to investigate the MIMO antenna performance under bending condition (Fig. 18). MIMO antenna is bent in two different orientations i.e. along x-and y-axis and the corresponding simulated S_{nn} parameters are shown in Figs. 19 and 20, respectively.

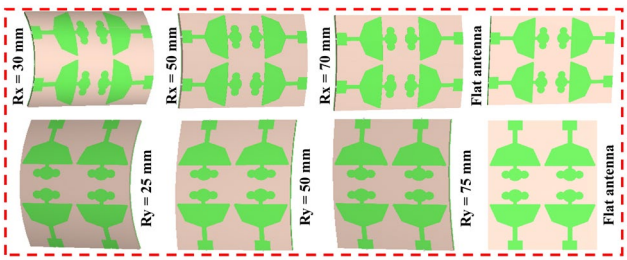


FIGURE 18. Deformation of antenna on multi-layer phantom along x-and y-orientations.

It can be seen that both the resonance frequencies slightly shift towards lower side when the deformation is maximum i.e. Rx = 30 mm (Fig. 19). The corresponding measured S-parameter is also compared at bending radius of 30 mm and found in good agreement with the simulated result. When antenna is bent along y-axis, there is negligible shift in simulated lower resonance frequency, however, the shift in higher resonance frequency is maximum at Ry = 25 mm (Fig. 20). Thus, from the results, it can be concluded that the proposed MIMO antenna is suitable for flexible/wearable electronics [30].

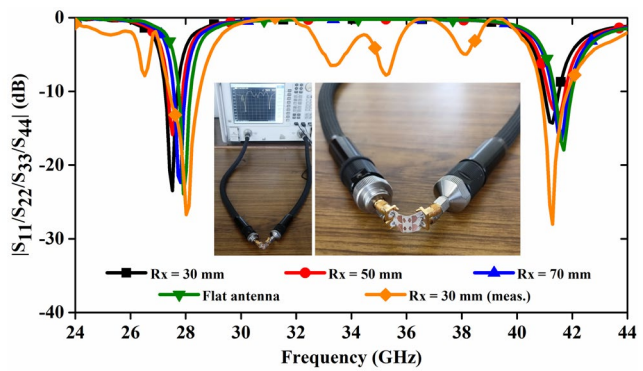


FIGURE 19. Measurement of S_{nn} - parameter against bending radii along x-axis.

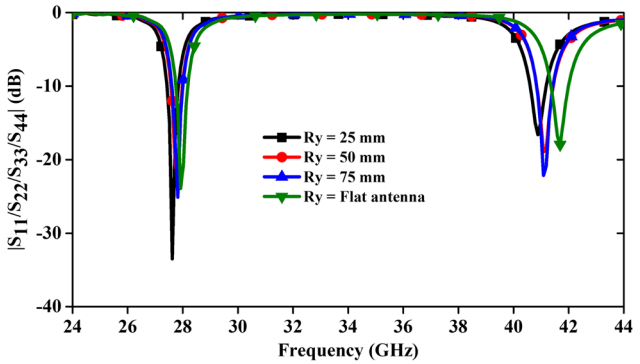


FIGURE 20. Simulated Snn - parameter variation for different bending radii along y-axis.

VIII. LINK BUDGET ANALYSIS

The proposed mm-wave MIMO antenna is designed for high-speed wireless body area network (WBAN). Therefore, it is required to calculate the link budget parameters between the transmitter and receiver (on-body).

TABLE I
LINK BUDGET PARAMETERS

Transmitter		
	Frequency (GHz)	28/41.69
G_t	Antenna gain (dBi)	6.04/5.93
P_t	Transmitter power (dBm)	16
	EIRP (dBm)	22.04/21.93
Receiver		
G_r	Receiver antenna gain (dBi)	4.1
T_o	Ambient temperature (K)	293
k	Boltzmann constant	1.38×10^{-23}
N_o	Noise power density (dB/Hz)	-203.9
Signal quality		
B_r	Bit rate (Mbps)	1, 20, 50, 100
E_b/N_o	Ideal PSK (dB)	9.6
G_c	Coding gain (dB)	0
G_d	Fixing deterioration (dB)	2.5

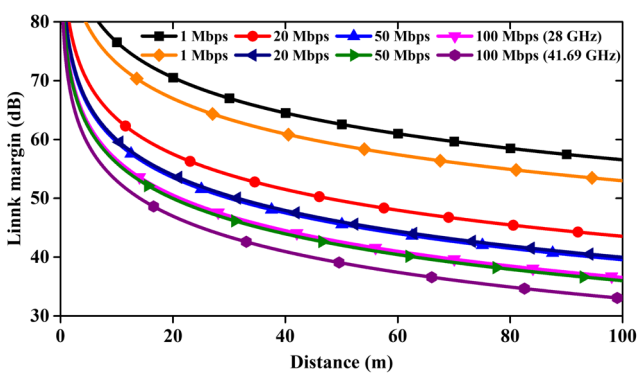


FIGURE 21. Calculated link margin with distance between transmit (T_x) and receiver (R_x) antenna.

The link margin (LM) is calculated using [31, 32] at two resonance frequencies 28/41.69 GHz and the link budget

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3412712

parameters are given in Table I. Here, we have used the input power of 500 mW to avoid the interference/noise from the other working devices at the same frequency band. The calculated LM is depicted in Fig. 21. From this figure, it is clear that the MIMO antenna achieves a bit rate of 100 Mbps with LM > 33 dB at 41.69 GHz frequency (upto 100m distance). Hence, the designed antenna is suitable for WBAN at two reported frequencies (28/41.69 GHz).

IX. ON BODY MEASUREMENT OF MIMO ANTENNA AND SAR (SPECEFIC ABSORPTION RATIO) Calculation

The proposed mm-wave MIMO antenna can be installed on the human body due to its flexible nature [33]. The measurement setup for the proposed MIMO design on human chest and hand are depicted in Fig. 22. The S-parameters of the antenna on chest and hand positions are presented in Fig. 23. From this figure, it can be observed that both the resonance frequencies are slightly altered and found at 27.91 and 40.54 GHz when antenna is located on hand. However, the shift of resonance frequencies is found slightly more when antenna is placed on the chest. The shift of resonance frequencies may be due to more complex anisotropic behaviour of tissues, bones and muscles on the chest portion of the body.

FIGURE 22. On body measurement of the MIMO antenna.

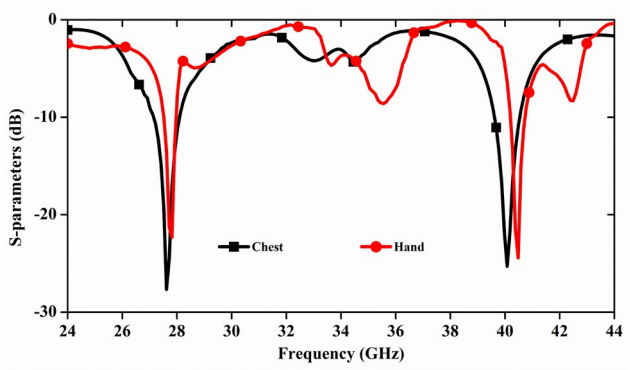


FIGURE 23. Measured S-parameters on human body (chest and hand).

Furthermore, to ensure the appropriate use of the antenna for on body applications, electromagnetic radiation emitted from the antenna should be analyzed. According to the International Commission on Non-Ionizing Radiation Protection (ICNIRP) and IEEE C95.1-2019 standard [34], for on-body application, the average SAR value should be below 1.6 and 2.0 W/Kg (for 1 gm and 10 gm of tissues). The SAR values are evaluated at two different frequencies (at 28 GHz and 41.69 GHz) in the multi-layer of human body phantom. The size of the phantom is 50×50 mm^2 with thickness of muscle, fat and skin of 20, 5 and 4 mm, respectively. To evaluate the SAR value, the electrical parameters for the different layers of human body are taken from [35].

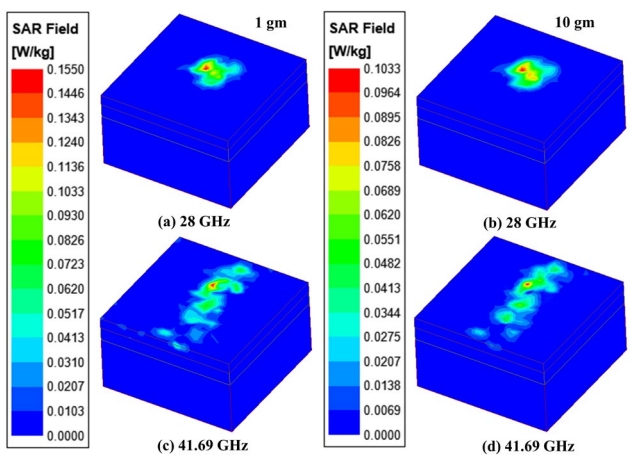


FIGURE 24. SAR calculation at 28 GHz, and 41.69 GHz.

Fig. 24 shows the 1gm/10gm average SAR values at 28/41.69 GHz (for input power of 500 mW) are 0.16/0.10 W/kg and 0.13/0.11 W/kg, respectively. This MIMO antenna is finally compared with recently reported antennas (Table II). None of the design has reported the bending analysis, CC, link margin and SAR analysis which shows the state of the art work in the proposed MIMO antenna.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3412712

TABLE II
COMPARISON OF PROPOSED ANTENNA WITH PAST REPORTED ANTENNAS

Ref.	Port	Volumetric size (mm^3)	Bandwidth (GHz)	Isolation (dB)	Gain (dBi)	ECC	CCL (bits/s/Hz	Measured Bending Analysis	CC (bps/Hz)	LM (dB)	SAR (1/10 gm)
[5] (2022)	2	11.4×5.3×0.8 (48.34)	29 (center frequency)	>36	6	<0.0001	NR	Yes	NR	NR	NR
[6] (2021)	4	32.5×32.5×0.51 (538.69)	36.83-40.0	>45	6.5	0.001	0.6	NO	NR	NR	NR
[7] (2022)	2	7.5×8.8×0.25 (16.50)	28/38 (center frequency)	>20	6.6	Value not given	NR	NO	NR	NR	NR
[9] (2023)	4	48×12×0.254 (146.30)	23-33/ 37.75-41	>20	5.7	0.00015	NR	NO	NR	NR	NR
[10] (2023)	4	20×20×0.20 (81.2)	26.9-29/ 34.5-41	>24	9.5	0.0001	≤0.4	NO	NR	NR	NR
[11] (2024)	2	18×9.2×0.787 (130.33)	24-28.8/ 36.6-40.8	>30	7.8/6	<0.0001	NR	NO	NR	NR	NR
[12] (2023)	2	26×14×0.762 (277.37)	27.1-28.8/ 35.2-38.9	>34.6	6.5/5.7	0.2×10^{-4}	NR	NO	NR	NR	NR
[13] (2024)	4	9.8×38.64×0.508 (189.34)	36.5-39.28	>30	10	3×10^{-5}	NR	NO	NR	NR	NR
[14] (2023)	4	40×72×0.508 (146.30)	24.5/28.5 (center frequency)	>33	9.8/8.7	0.0001	NR	NO	NR	NR	NR
[15] (2022)	2	24×15×0.787 (283.32)	Not reported	>26	5.58	NA	NR	NO	NR	NR	NR
[16] (2022)	3	26×26×0.76 (513.76)	27.5-28.05	>30	4.71	<0.002	NR	NO	NR	NR	NR
Proposed work	4	16×12×0.25 (48.00)	27.75-28.18/ 41.31-41.99	>20	6.05/5.78	≤0.10	≤0.28	Yes	20.86	>33 (upto 100 m)	(0.16/0.10) and (0.13/0.11)

Note: NR = Not Reported

X. CONCLUSION

A dual-band 4×4 MIMO flexible antenna for millimeter wave applications has been designed, fabricated and measured. Theoretical resonance formula is developed for the proposed radiating structure to design the antenna at specific resonance frequency. The isolation >20 dB is achieved without any decoupling structure. The ECC, DG, MEG, CCL, TARC, CC and CDF of the antenna ensure its candidature for the applications in high-speed communication systems. This antenna demonstrates the transmit/receive data rate of 100 Mbps upto a communication distance of 100 meter at both the resonance frequencies of 28/41.69 GHz. The MIMO antenna is measured on different parts of human body to assert its on-body applications. The SAR values are also calculated on multilayer phantom model and found quite below as compared to the predefined values at 1 and 10 g of tissues. Thus, the proposed simple, ultrathin and flexibile MIMO antenna is suitable for mm-wave 5G communications and wireless wearable electronics.

REFERENCES

[1] J. S. Park et al., "A tilted combined beam antenna for 5G communications using a 28 GHz band," *IEEE Antennas Wireless Propag. Lett.*, vol. 15, pp. 1685-1688, 2016.

[2] P. B. G. Shariff, P. R. Mane, P. Kumar, T. Ali, M. G. N.Alsath, "Planar MIMO antenna for mmWave applications: Evolution, present status & future scope," *Heliyon*, vol. 9, no. 2, e13362, 2023.

[3] M. N. Hasan, and M. Seo, "Compact omnidirectional 28 GHz 2×2 MIMO antenna array for 5G communications," Int. Symp. Antennas Propag. (ISAP), Korea, 2018.

[4] P. Sharma, R. N. Tiwari, P. Singh, P. Kumar, and B. K. Kanaujia, "MIMO antennas: Design approaches, techniques and applications," *Sensors*, vol. 22, no. 20, Article no. 7813, 2022.

[5] A. Ahmad, D. Choi, and S. Ullah, "A compact two elements MIMO antenna for 5G communication," *Sci. Rep.*, vol. 12, Art. no. 3608, 2022.

[6] D. A. Sehrai et al., "Compact quad-element high-isolation wideband MIMO antenna for mm-wave applications," *Electronics*, vol. 10, no. 11, Art. no. 1300, 2021.

[7] A. E. Farahat, and K. F. A. Hussein, "Dual-band (28/38 GHz) wideband MIMO antenna for 5G mobile applications," *IEEE Access*, vol. 10, pp. 32213-32223, 2022.

[8] A. R. Sabek, A. A. Ibrahim, and W. A. Ali, "Dual-band millimeter wave microstrip patch antenna with stub resonators for 28/38 GHz applications," J. Phys.: Conf. Ser. 2128, 012006, 2021.

[9] M. E. Munir et al., "A four element mm-wave MIMO antenna system with wide-band and high isolation characteristics for 5G applications," *Micromachines*, vol. 14, no. 4, Art. no. 776, 2023.

[10] N. Sghaier et. Al., "Millimeter-wave dual-band MIMO antennas for 5G wireless applications," *J. Infrared, Millimeter, Terahertz Waves*, vol. 44, pp. 297-312, 2023.

[11] D. Khan, A. Ahmad, and D. Y. Choi "Dual-band 5G MIMO antenna with enhanced coupling reduction using metamaterials," *Sci. Rep.*, 14, Article no. 96, 2024.

[12] B. A. F. Esmail, and S. Koziel, "Design and optimization of metamaterial-based dual-band 28/38 GHz 5G MIMO antenna with

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3412712

modified ground for isolation and bandwidth improvement," *IEEE Antennas Wireless Propag. Lett.*, vol. 22, no. 5, pp. 1069-1073, 2023.
[13] A. K. Singh, and S. Pal, "Compact self-isolated extremely low ECC folded-SIW-based slot MIMO antenna for 5G application," *IEEE Antennas Wireless Propag. Lett.*, vol. 23, no. 1, pp. 194-198, 2024.
[14] M. Aboualalaa, and I. Mansour, "Dual-band end-fire four-element MIMO antenna array using split-ring structure for mm-wave 5G applications," *IEEE Access*, vol. 11, pp. 57383-57390, 2023.
[15] M. M. El-Wazzan, H. H. Ghouz, S. K. El-Diasty, and M. A. A. Dahab, "Compact and integrated microstrip antenna modules for mm-wave and microwave bands applications," *IEEE Access*, vol. 10, pp. 70724-70726, 2022.
[16] E. Kobal, R. J. Liu, C. Yu, and A. Zhu, "A high isolation, low-profile, triple-port SIW based annular slot antenna for millimeter-wave 5G MIMO applications," *IEEE Access*, vol 10, pp. 89458-89464, 2022.
[17] J. P. Xu, L. J. Xu, L. Ge, Z. Duan, and Y. M. Tang, "Flexible wearable antenna based on MIMO technology," IEEE MTT-S Int. Wireless Sympos. (IWS), pp 1-3, 2019.
[18] J. Mayer, L. Matter, B. Nuss, J. Kowalewski, and T. Zwick, "Conformal antennas for a wide view angle in automotive radar," in Proc. Ger. Microw. Conf. (GeMiC), pp. 119-122, 2019.
[19] A. Kumar, P. Pattanayak, R. K. Verma, D. Sabat, and G. Prasad, "Two-element MIMO antenna system for multiband millimeter-wave, 5G mobile communication, Ka-band, and future 6G applications with SAR analysis," *AEU-International Journal of Electronics and Communications*, vol. 171, Art. No. 154876, 2023.
[20] A. A. Althuwayb et al., "Metasurface-inspired flexible wearable MIMO antenna array for wireless body area network applications and biomedical telemetry devices," *IEEE Access*, vol. 11, pp. 1039-1056, 2023.
[21] A. K. Biswas, S. Biswas, S. Haldar, and A. Nanndi, "A highly decoupled flexible 4-element MIMO antenna with band notched characteristics for ultra wide-band wearable applications," *AEU-International Journal of Electronics and Communications*, vol. 173, Art. No. 154985, 2024.
[22] U. Ullah, S. Koziel, and A. P. Dabrowska "Design and characterization of a planar structure wideband millimeter-wave antenna with wide beamwidth for wearable off-body communication applications," *IEEE Antennas Wireless Propag. Lett.*, vol. 21, no. 10, pp. 2070-2074, 2022.
[23] K. P. Ray, M. D. Pandey, and S. Krishnan "Determination of resonance frequency of hexagonal and half hexagonal microstrip antenna," *Microwave Opt. Technol. Lett.*, vol. 49, no. 11, pp. 2876-2879, 2007.
[24] I. J. Bahl, and P. Bhartia, "Microstrip antennas," Artech House, Boston, Dedham, MA, 1980.
[25] R. N. Tiwari, P. Singh, P. B. Barman, "Triple-band gap coupled 4 × 4 MIMO antenna in mm-wave for high data rate and IoT applications," Emergent Converging Technologies and Biomedical Systems (ETBS 2023), Lecture Notes in Electrical Engineering, vol 1116. Springer, Singapore, 2024.
[26] T. Taga, "Analysis for mean effective gain of mobile antennas in land mobile radio environments," *IEEE Trans. Veh. Technol.*, vol. 39, no. 2, pp. 117-131, 1990.
[27] H. Shin, and J. H. Lee, "Capacity of multiple-antenna fading channels: spatial fading correlation, double scattering, and keyhole," *IEEE Trans. Inform. Theory*, vol 49, no. 10, pp. 2636-2647, 2003.
[28] M. Manteghi, and Y. R. Samii, "Multiport characteristics of a wideband cavity backed annular patch antenna for multipolarization operations," *IEEE Trans. Antennas Propag.*, vol. 53, no. 1, pp. 466–74, 2005.
[29] Z. Li, Z. Du, M. Takahashi, K. Saito and K. Ito, "Reducing mutual coupling of MIMO antennas with parasitic elemets for mobile terminals," *IEEE Antennas Wireless Propag. Lett.*, vol. 60, no. 2, pp. 473-481, 2012.
[30] H. Lee and Y. B. Park, "Wideband ring-monopole flexible antenna with stub for WLAN/C-band/X-band applications," *Appl. Sci.*, vol. 12, no. 21, 10721, 2022.
[31] H. Rajagopalan, Y. R. Samii, "Wireless medical telemetry characterization for ingestible caplsule antenna designs," *IEEE Antennas Wireless Propag. Lett.*, vol. 11, pp. 1679-1682, 2012.
[32] R. N. Tiwari, V. Kaim, P. Singh, T. Khan, and B. K. Kanaujia, "Semi-flexible diversified circularly polarized millimeter-wave MIMO antenna for wearable biotechnologies," *IEEE Trans. Antennas and Propag.*, vol. 71, no. 5, pp. 3968-3982, 2023.
[33] M. Särestöniemi, M. Sonkki, S. Myllymäki, and C. Pomalaza-Raez, "Wearable flexible antenna for UWB on-body and implant communications," *Telecom.*, vol. 2, no. 3, pp. 285-301, 2021.
[34] Conference, I. EMF, New ICNIRP Guidelines and IEEE C95.1-2019 Standard: Differences and Similarities. 5, 1-30(2019).
[35] https://itis.swiss/virtual-population/tissue properties/database/ dielectric-properties.

Rakesh Nath Tiwari received the B.Sc. degree (electronics and physics) from the University of Allahabad, Allahabad, India, in 2002, the M.Sc. degree (electronics) from Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur, India, in 2004, the M.Tech. degree in optical and wireless communication technology with Gold Medal from the Jaypee University of Information Technology, Solan, India, in 2008, and the Ph.D. degree from the Department of Electronics and Communication Engineering, Uttarakhand Technical University, Dehradun, India, in August 2020. He is currently working as Assistant Professor with the Department of Electronics and Communication Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India.

Dr. Tiwari has published one patent, eighteen research papers in SCI journals, four Scopus journals, three book chapters, and over fifteen articles in international conferences, and his article have been cited over 550 times with h-index of 12. He is the recipient of the Best Paper Award from AEU-International Journal of Electronics and Communication, Elsevier 2020. His current research interests include design and modeling of slot patch antennas, UWB antennas, circularly polarized microstrip antennas, MIMO antennas, wearable antennas, capsule endoscopy, and millimeter-wave applications. He is a life member of Materials Research Society of India (MRSI), and International Association of Engineers (IAENG). He is a reviewer for many reputed journals.

Deepti Sharma received the B.Tech. degree from Uttar Pradesh Technical University, Lucknow, India, in 2005, MTech. degree from Guru Gobind Singh Indraprastha University, Delhi, India, in 2011, and Ph. D from Jawaharlal Nehru University, New Delhi, India in 2023. She is currently working as an Assistant Professor with the Department of Electronics and Communication Engineering, G.L. Bajaj Institute of Technology and Management, Greater Noida, Uttar Pradesh, India. She has published 07 patents and 07 research articles in several peer-reviewed international journals. Her current research interests include MIMO antennas and implantable and wearable antennas for biomedical applications.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3412712

Prabhakar Singh was born in village Semara, Chandauli (U.P), India in 1984. He received his B.Sc. and M.Sc. degree in Physics from V. B. S. Purvanchal University in 2004 and 2006, respectively. He received his Ph.D. degree from J. K. Institute of Applied Physics, Department of Electronics and Communication, University of Allahabad, India in 2010. During the Ph.D. work, he also worked in the Department of Electronics Engineering, ITT BHU. He taught B. Tech. and M. Tech. students at Delhi Technological University, Delhi, India for one year from 2009 to 2010. He worked as assistant professor at Bahra University, Shimla Hills, Himanchal Pradesh, India from 2010 to 2011. Presently, he is working as Professor and Head of the Physics Department, Galgotias University, Greater Noida, India. He has published more than 70 research papers in peer reviewed International/National journals and conference proceedings with more than 1100 citations with h-index of 20. He has published 3 Indian Patents. He is reviewer of many International and National Journals which includes IEEE Access, Electronic Letters, wireless communication, sensors etc. He is presently working on broadband microstrip antennas, size miniaturization techniques in patch antennas, UWB, RF harvesting, MIMO designs, flexible and implantable antennas.

Pradeep Kumar obtained the bachelor's, master's, and PhD degrees in electronics and communication engineering, in 2003, 2005, and 2009, respectively. He was a postdoctoral researcher with the Autonoma University of Madrid, Spain. He has over 18 years of experience in academics and research. He has held various positions, such as a lecturer, a senior lecturer, an assistant professor, and an associate professor. He is currently working with the University of KwaZulu-Natal, South Africa. He is the author of many research papers published in various peer-reviewed journals/conferences. He is registered as a Professional Engineer with the Engineering Council of South Africa. He is serving as a Reviewer/a TPC Member for many journals/conferences. His current research interests include antennas, antenna arrays, and wireless communications.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

JMEPEG
https://doi.org/10.1007/s11665-024-09701-2

©ASM International
1059-9495/$19.00

ORIGINAL RESEARCH ARTICLE

Determination of Residual Stresses and Surface Roughness Using Chemical Assisted Ball End Magnetorheological Finishing

Anand Sharma and Mahendra Singh Niranjan

Submitted: 14 March 2023 / Revised: 27 May 2024 / Accepted: 2 June 2024

The generation of residual stresses by grinding process on the surface for a soft material results in poor performance and diminished life. Ball end magnetorheological finishing (BEMRF) process is a recently developed process that is effectively used for fine figuring and polishing of a variety of magnetic and non-magnetic materials. In this paper, grinding-induced residual stresses are addressed, and an attempt is made to relieve these residual stresses while achieving high surface finish using chemical assisted ball end magnetorheological finishing (CA-BEMRF) process. The impact of various CA-BEMRF variables on percentage surface roughness reduction and percentage reduction in residual stresses are discussed statistically and graphically. Residual stresses of workpiece surfaces have been measured by using portable x-ray residual stress analyzer. Scanning electron microscopy (SEM) of the finished workpiece before and after CA-BEMRF process is carried out to assess the effects of magnetizing current, tool rotation and working gap on surface topography of machined surface. It has been observed that reduction in residual stress is achieved along with high surface finish on aluminum workpiece surface using CA-BEMRF technique. Significant process parameters affecting the residual stresses and surface roughness during polishing of workpiece are obtained using analysis of variance (ANOVA) and *F*-test. The optimization problem for maximum percentage reduction in surface roughness and maximum percentage reduction in residual stress is formulated as multiobjective, multivariable, nonlinear optimization problem. Maximum percentage surface roughness reduction and minimum residual stress has been found at magnetizing current 2.3A, rotational speed of tool 550 rpm and working gap of 0.5 mm. Confirmatory experimental tests have been conducted and results obtained were found very close to the predicted outcome.

Keywords aluminum, ANOVA, chemical assisted ball end magnetorheological, residual stress, SEM

1. Introduction

There is a significant demand for nano-meter and macro-meter items possessing excellent polished surface and strength in this age of new industrialization revolution. The surface finish has a direct impact on the optical quality, mechanical characteristics, and bio-medical functions of microparts (Ref 1, 2). Engineering materials are utilized in a variety of applications to meet industrial objectives by leveraging qualities like conductivity, ductility, and lightweight (Ref 3). However, some of these characteristics, such as malleability and weak scratch resistance, also provide challenges when polishing metals (Ref 4). For the finishing of copper alloy at the nanometric level, numerous successful attempts have been made (Ref 5, Ref 6). Unfortunately, relatively few research has focused on using chemo-mechanical polishing (CMP) and magnetorheological finishing to achieve nanometric-level finishing on aluminum alloys (MRF). Optical mirror coatings have frequently used silver and gold. Today's high-quality mirrors, however, are coated with aluminum rather than silver to prevent tarnishing caused by air exposure, particularly sulfur, which results in the development of brown colored tarnish spots. These spots also result in a decrease in reflectance and a rise in emissivity (Ref 7). Gold is a precious metal that is often used in jewelry and other decorative items. However, it is very costly as compared to other metals such as aluminum. Therefore, aluminum is often preferred over gold for various purposes. It is lighter in weight, cheaper, and more durable than gold. Additionally, it is easier to work with, making it a popular choice for many applications.

Humans' quest for solar and space exploration has led to a need for nanofinished aluminum alloys (Ref 8). Telescope metal mirrors made of aluminum alloys are completed using the DTM process (Ref 9, 10). These mirrors are thin, highly conductive, very resistant to corrosion, and have a longer fatigue life. Other significant uses of nanofinished aluminum surfaces include diaphragms for optical sensors (Ref 11), very corrosion-resistant marine transportation, microelectronic-based devices (Ref 12), and mass level integration for

This invited article is part of a special topical issue of the Journal of Materials Engineering and Performance on Residual Stress Analysis: Measurement, Effects, and Control. The issue was organized by Rajan Bhambroo, Tenneco, Inc.; Lesley Frame, University of Connecticut; Andrew Payzant, Oak Ridge National Laboratory; and James Pineault, Proto Manufacturing on behalf of the ASM Residual Stress Technical Committee.

Anand Sharma and **Mahendra Singh Niranjan**, Delhi Technological University, New Delhi, Delhi, India. Contact e-mail: doc.asharma@yahoo.in.

Published online: 14 June 2024

conductive wires (Ref 13). It has been discovered that finishing aluminum is a difficult process due to its malleability and flexibility. Nowadays, due to a few drawbacks including burrs, high residual stresses, sub-surface damages, etc., it is not viable to polish soft metals like aluminum up to the nanometric level using standard finishing methods like grinding, sprucing, burnishing, and lapping.

Residual stresses are stresses that remain in a material after it has undergone some process. These stresses can have harmful effects on the material, causing it to break or fail. Residual stresses can be caused by many factors, such as welding, casting, machining, or even just from the materials used in the construction of the item. The effects of residual stresses can be minimized by careful design and engineering of the item, as well as by using materials that are less likely to be affected by them.

To achieve high surface finish and minimum residual stress on aluminum alloys few finishing approaches which were able to control the finishing forces exerted on the workpiece surface are magnetic flow polishing, machining through abrasive flow, magnetorheological-based abrasive flow and jet flow. In those approaches, the finishing forces are managed via way of means of controlling the magnetic flux density by means of the use of permanent magnet or an electromagnet. MRF process have been very successful in finishing of various materials such as glass, molds made of steel, and EN-31, polyamide but very few research has been done on aluminum alloys using this technique. However, these techniques can finish limited geometric shapes as concave, convex, planar, and symmetrical spheres.

Inspired by the favorable features of chemical mechanical polishing (CMP) process, various researchers have made an attempt to hybrid the CMP and MRF process to obtain high surface finish for different materials. Jain et al. (Ref 14) developed chemo-mechanical MR finishing process by associating CMP and MRF process to enhance the surface finish of silicon blanks. It was concluded that operational gap, finishing period and rotational speed of magnets had significant effect and led to a surface finish of 4.8 Å.

Ranjan et al. (Ref 5, 15) super finished a copper alloy using CMMRF technique and developed a mathematical model to study the polishing pressure and other controlling parameters. A delicate chemically passivated layer was created on copper alloy surface by chemically reacting it with ammonia, glycine and nitric acid. The results showed that optimization of pH value of the chemical is important to achieve better surface finish. Results showed that the depth and width of scratches on the surface were a few tens of nanometers size.

Ranjan et al. (Ref 16) theoretically explored magnetism, polishing pad formation and polishing pressure using FEA-CFD simulation, the results of which were validated later experimentally on aluminum 6061-T6 alloy. The results revealed that the polishing pressure is affected by three parameters, viz. Working gap, MR fluid and rotational speed.

Ghai et al. (Ref 17) achieved a nano-level surface finish (0.597 nm) on Al 7075-T6 alloy using CMMRF technique. These results were obtained by optimizing the concentration of chemicals, abrasives, and CIP particles along with working gap. The authors used deionized water, glycine, ammonia, and citric acid to form a brittle passivated layer on workpiece surface and model the experiments with central composite design model technique. It was observed that aluminum alloy could only be finished in alkaline solution.

Kumar and Singh (Ref 18) critically assessed the benefits, futuristic opportunities, and the challenges in improving the capabilities of CMMRF process. They beautifully explained the mechanism and various process parameters affecting the capability of CMMRF process and its applications in the field of aerospace research, solar, medical implants, electronics and optical. It was also discussed that in future more focus can be given to in-depth finishing of ferrous and non-ferrous material, brittle materials, complex structures, etc. It was concluded that this technique can produce a surface finish on both brittle and ductile materials in the range of 2-6 Å.

Although various researchers have carried out the finishing of materials like silicon, copper, aluminum alloy using "CMMRF" process, but no work was found via BEMRF technique. In addition, according to the past work very limited work has been carried out using aluminum 7075 alloy. The present work is aimed to achieve a high surface finish on aluminum 7075 alloy flat workpiece which can be directly used in different applications. The approach described in this study uses a newly created process called chemical assisted ball end magnetorheological finishing (CA-BEMRF), which combines the CMP and BEMRF procedures (Ref 19, 20), to finish the surface of an aluminum alloy workpiece. To see how a chemical reaction affects the residual stresses and surface finish of an aluminum alloy workpiece, the percentage reduction in residual stress and percentage surface roughness reduction obtained before and after the CA-BEMRF process were compared (Ref 21-24).

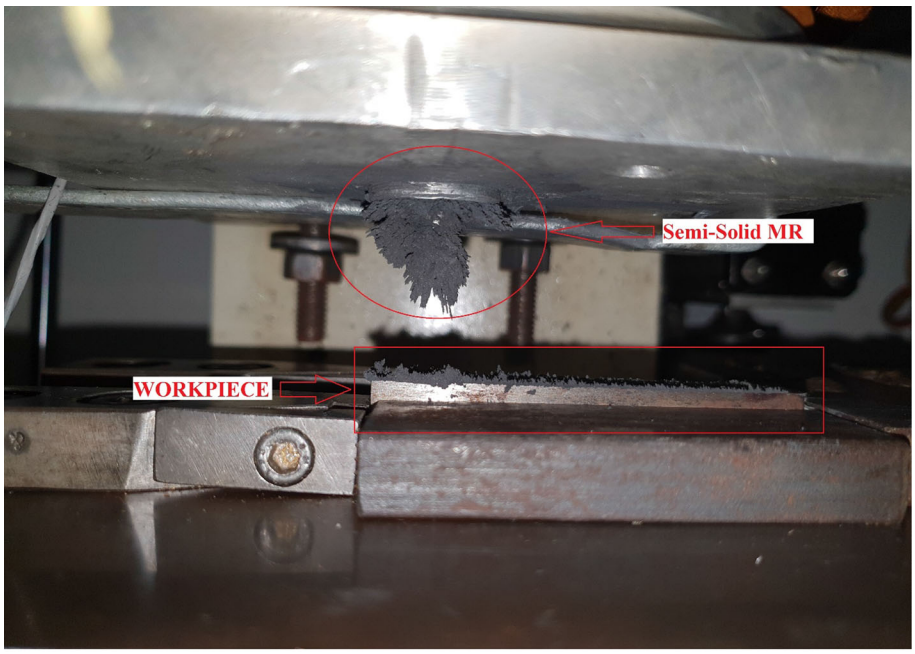
The BEMRF process is a novel way of reducing residual stress in materials. By passing an electric current through the material, the BEMRF process can help to remove any internal stresses that may be present. This can be beneficial for a number of reasons, including improving the overall strength and durability of the material. Additionally, the BEMRF process can also help to improve the dimensional accuracy of the material, by reducing any warping or distortion that may be caused by residual stress.

2. Magnetorheological Fluid

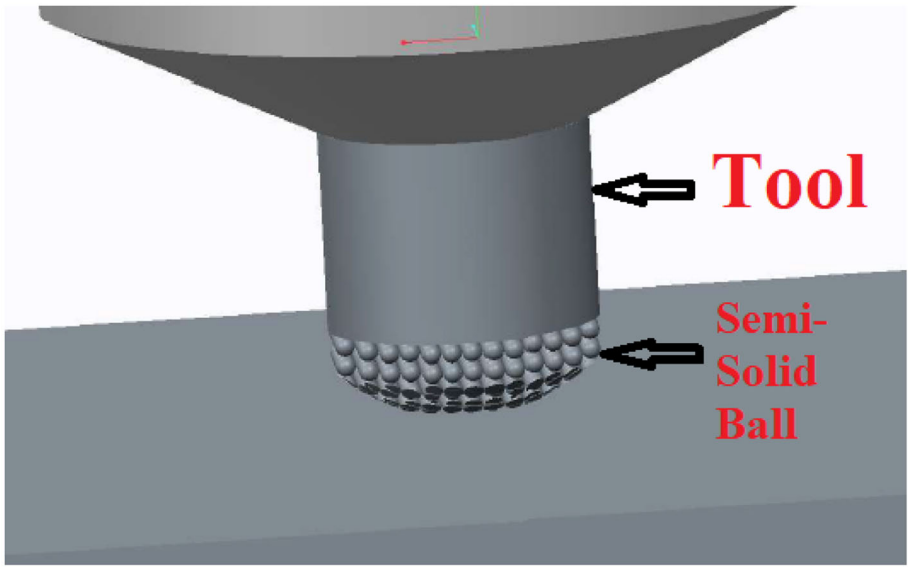
In CMP, MRF, CMMRF, and BEMRF processes, magnetorheological fluid (MR fluid) is frequently used to polish surfaces to an ultra-smooth finish. The preparation of the MR fluid is the most important factor in completing any surface in the CA-BEMRF. MR fluid enables the control of finishing forces applied on the workpiece surface while finishing through CA-BEMRF process. Table 1 describes the key constituents of MR fluid prepared for the experimentation. MR fluid consists of base fluid, abrasive particles (SiC) and carbonyl iron particles (CIPs) to form semi-solid ball under the effect of magnetic field as shown in Fig. 1(a). The schematic of the formation of ball-shaped MR fluid is shown in Fig. 1(b). The MR fluid stiffens and attains the shape of a polishing ball under the action of the magnetic current before being compressed against the workpiece surface. Because the magnetic force increases as the working gap decreases, ultra-finishing can be accomplished by adjusting the distance between the tool tip and the workpiece. Controlling the tool's rotational speed, working gap, and magnetizing current the finishing forces can be controlled.

Table 1 Constituents of magnetorheological fluid

S. No.	Constituents of MR fluid	Composition	W/W%
1	Base fluid	Paraffin oil (80 wt%) and grease (20 wt%)	55
2	Abrasive	Silicon carbide (SiC)	25
3	Magnetic particles	Carbonyl iron particles	20



(a)



(b)

Fig. 1 (a) Semi-solid MR fluid ball under influence of magnetic current, (b) schematic of semi-solid ball

3. Mechanism of Finishing

Finishing through CA-BEMRF process is accomplished through a mix of chemical action and mechanical erosion. Under the influence of a magnetizing current, magnetic particles create a semi-solid ball with abrasives, which causes mechanical material removal. A chemical agent then causes the production of a passivated layer, which is removed by the magnetic and abrasive particles.

3.1 Chemical Reaction

Chemical reaction is essential for forming passivated layer to protect the aluminum alloy. A passive layer on the surface of aluminum alloy is formed when the workpiece is dipped in the chemical solution consisting of H_2O_2. Citric acid and phosphoric acid (H_3PO_4) were used as pH buffer solutions. In this process, hydrogen peroxide (H_2O_2) was used as an oxidant for softening the aluminum alloy workpiece surface which in turn increases the surface finish after mechanical action through BEMRF process. Kuo and Tsai (19) showed that a small concentration of H_2O_2 can substantially improve the material removal rate which in turn can quite effectively increase the surface finish of the workpiece. Equation 1, 2, and 3 outline the mechanism's reaction pathway.

$$2Al + 3H_2O_2 \rightarrow Al_2O_3 + 3H_2O \rightarrow 2Al(OH)_3 \quad \text{(Eq 1)}$$

$$Al(OH)_3 + 3H_2O \rightarrow 2Al(OH_2)_3(OH)_3 \quad \text{(Eq 2)}$$

$$Al(OH_2)_3(OH)_3 + H_3PO_4 \rightarrow AlPO_4 + 6H_2O \quad \text{(Eq 3)}$$

This mechanism states that the surface of the Al is first oxidized by H_2O_2 to produce Al_2O_3. Al_2O_3 is next hydrated and reduced (dissolved) by H_3PO_4 to create a water-soluble phosphate salt (i.e., $AlPO_4$), which is then mechanically removed by CI and abrasive particles, concluding the removal cycle.

3.2 Mechanical Action

Using an electromagnet, magnetic field is generated in CA-BEMRF process which shapes MR fluid into a semi-solid ball. Silicon carbide abrasive particles are held in chains of magnetic carbonyl iron particles when the magnetizing current is switched ON. The flexible finishing pad developed a micro-texture due to the average CIPs size of 1 to 5 μm. As seen in Fig. 2(b), microtextures aid in retaining SiC abrasive particles and transferring finishing force to abrasive particles. By applying magnetic force to abrasive particles via CIPs, a passive layer that had formed on the aluminum workpiece surface was removed, and peak levels were reduced to a certain extent. CIPs hold the abrasive particles while also smoothing the workpiece's surface and removing any peaks. Tinier magnetic and soft abrasive particles are usually advised because hard abrasives and substantial magnetic particles will cause scratches. The final finishing force (FF) is a resultant of the normal force (N_F) and shear force (SH_F), as illustrated in Fig. 2(b), based on the abrasive and magnetic particle microstructure. The shear force is responsible for material removal, whereas *Fn* is in duty of determining the depth of penetration of the workpiece surface.

4. Experimental Setup

As a result of the restriction on the relative mobility of the workpiece and finishing tool, the limitations of finishing limited geometries were eliminated by developing ball end magnetorheological finishing (BEMRF), a variation of the MRF method. The original BEMRF configuration contained a MR finishing tool oriented vertically which is powered by a servo motor. The tool consists of a cylindrically formed central core (prepared of iron), an electromagnetic coil rapped around the tool & exterior core orientated concentrically to one another. The electromagnetic coil is intended to achieve around 0.8 Tesla flux density.

Figure 2(a) depicts the configuration of a 3-axis CNC BEMRF tool powered by a servo motor. The electromagnetic coil is enclosed by a tube made of copper through which cold water is pumped to keep the temperature acceptable for experiments. Once the current supply is turned on, the formation of a virtually very viscous MR fluid "ball-shaped" at the tool tip takes place as shown in Fig. 1(a). The viscosity of MR fluid can be controlled or changed by real-time control of the magnetizing current, thereby controlling the intensity of the magnetic field.

4.1 Preparation of Sample

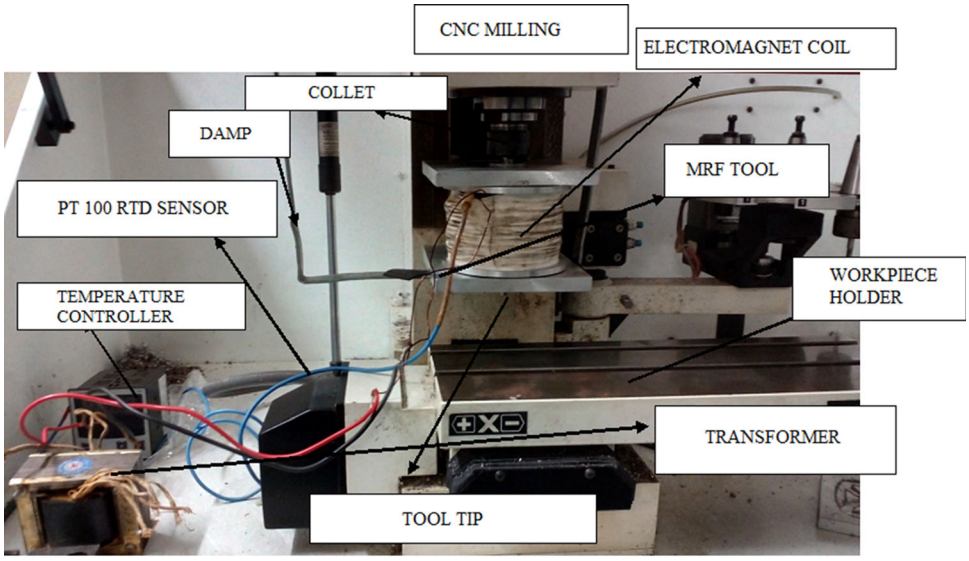
Workpieces of Al 7075 alloy, each measuring 70 mm × 10 mm × 10 mm was cut and pre-finished using precision grinding technique. Table 2 shows the grinding process parameters used for pre-finishing the workpieces. The specimens were then washed with acetone and rinsed with DI water to eliminate any debris or burrs that remained after grinding. The Mitutoyo Surftest SJ-301 equipment is used to measure the first stage surface roughness acquired after grinding.

4.2 Preliminary Experiments

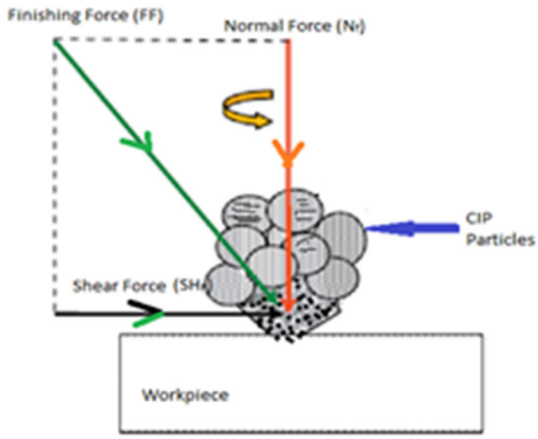
The preliminary experiments were performed on three-axis CNC BEMRF setup developed at Delhi Technological University, Delhi to obtain the ranges and significant process parameters. The effect of key parameters on percentage reduction in surface roughness and residual stress is investigated. The output response, i.e., percentage reduction in surface roughness (%RSR_C) and residual stress (%RRS) after CA-BEMRF process is calculated by Eq (4) and (5), respectively.

$$\%RSR_C = \frac{\text{Surface Roughness after grinding process} - \text{Surface Roughness after CA} - \text{BEMRF process}}{\text{Surface Roughness after grinding process}} \times 100 \quad \text{(Eq 4)}$$

$$\%RRS = \frac{\text{Residual Stress after grinding process} - \text{Residual Stress after BEMRF process}}{\text{Residual Stress after grinding process}} \times 100 \quad \text{(Eq 5)}$$



(a)



(b)

Fig. 2 (a) Setup of 3-axis BEMRF process and (b) abrasive and magnetic particle microstructure on the workpiece surface

Table 2 Precision grinding process parameters

Parameters	Values	Units
Wheel velocity, m/min	2026	m/min
Feed rate-work table traverse, m/min	0.12	m/min
Depth of cut	20	μm

Table 3 represents the preliminary experiments conducted on Al 7075 alloy workpiece surface using CA-BEMRF process. These experiments were conducted at 2.5A magnetizing current, 500 rpm rotational speed of tool and 1.5 mm working gap with variation in chemical composition ranging from 3 to 7w/w% of MR fluid. The specimen was dipped in the chemical solution for 30 minutes prior to finishing through MR fluid.

Table 3 Preliminary experiments with CA-BEMRF process

Exp no.	Chemical composition	Magnetizing current	Rotational speed of tool	Working gap	Initial Ra[nm]	Final Ra[nm]	%RSR$_C$
1	3	2.5	500	1.5	440	278	36.82
2	4	2.5	500	1.5	442	280	36.65
3	**5**	**2.5**	**500**	**1.5**	**410**	**255**	**37.81**
4	6	2.5	500	1.5	445	280	37.10
5	7	2.5	500	1.5	435	282	35.17

Bold values indicate the highest percentage reduction in surface roughness after CA-BEMRF process is obtained at these parameters

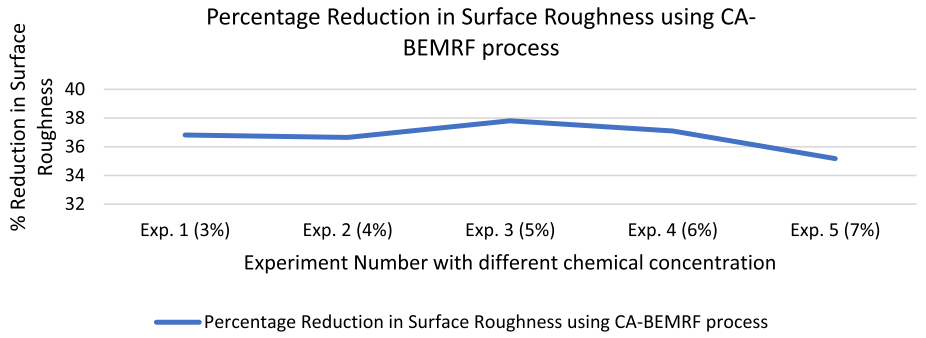


Fig. 3 Variation of percentage reduction in surface roughness using CA-BEMRF process

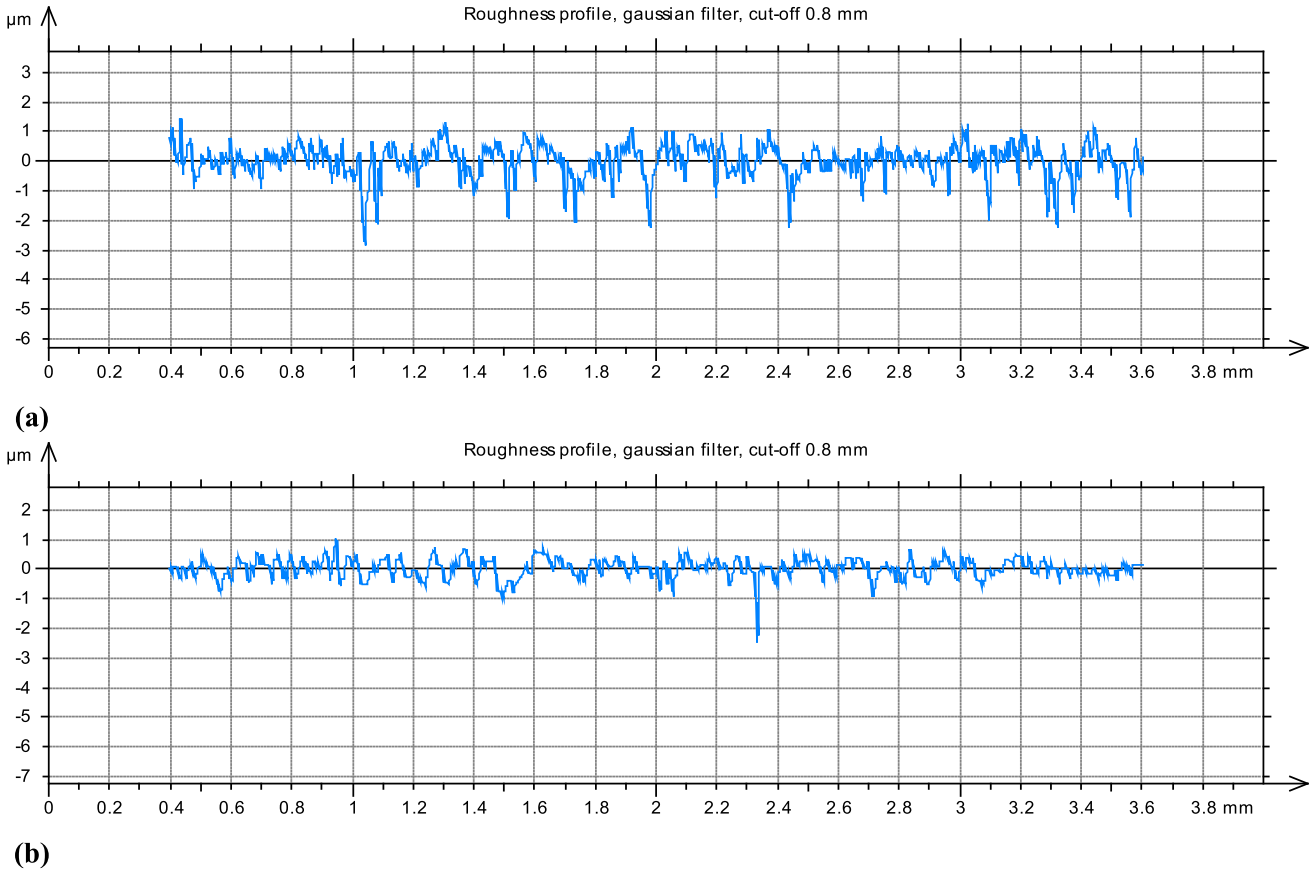


Fig. 4 (a) Roughness profile of Al 7075 workpiece surface after grinding (b) Roughness profile of Al 7075 workpiece surface after CA-BEMRF process at 5% chemical composition, 2.5A magnetizing current, 500 rpm rotational speed of tool and 1.5 mm working gap

It is observed from Fig. 3 that the highest $\%RSR_C$ is found to be 37.81% at 5% chemical composition, 2.5A magnetizing current, 500 rpm rotational speed of tool and 1.5 mm working gap using CA-BEMRF process. The roughness profile of Al 7075 alloy workpiece surface after grinding and after CA-BEMRF process is shown in Fig. 4(a) and (b), respectively. The surface roughness value of 410 nm obtained after grinding the workpiece surface was reduced to 255 nm using CA-BEMRF process.

Comparing the percentage reduction in surface roughness obtained after grinding process and the conduction of preliminary experiments using CA-BEMRF process, it is observed that the surface finish has improved using the assistance of chemical agent. Hence, 5 w/w% chemical composition of hydrogen peroxide was fixed for the detailed experiments.

5. Detailed Experiments

In this work, the experimentation for finishing of aluminum 7075 alloy workpiece is designed according to central composite design (CCD) technique using design of expert 6.0.7 software. A total of 20 experiments designed according to

Table 4 Levels and ranges of selected process parameters for CA-BEMRF process

		Levels						
		I	II	III	IV	V		
Process parameter	Units	− 2	− 1	0	1	2	Static parameters	Description
Magnetizing current (V_M)	A	1.5	2.0	2.5	3.0	3.5	Feed rate	50 mm/min
Rotational speed of tool (T_R)	rpm	300	400	500	600	700	Finishing time	30 mins
Working gap (W_X)	mm	0.5	1.0	1.5	2.0	2.5	Chemical composition	5 w/w%

Table 5 Percentage reduction in surface roughness and residual stress after CA-BEMRF process

Std order	Run order	V_M	T_R	W_X	Initial Ra	Final Ra	%RSR$_C$	Initial RS	Final RS	%RRS
6	1	1	− 1	1	427	254	40.56	65	21	67.69
14	2	0	0	2	433	315	27.32	63	34	46.03
17	3	0	0	0	417	262	37.1	78	29	62.82
10	4	2	0	0	427	255	40.32	87	28	67.82
3	5	− 1	1	− 1	440	232	47.32	72	20	72.22
16	6	0	0	0	404	255	36.86	53	13	75.47
13	7	0	0	− 2	442	118	57.32	46	10	78.26
8	8	1	1	1	439	311	29.12	85	55	35.29
1	9	− 1	− 1	− 1	315	211	32.97	75	28	62.67
20	10	0	0	0	316	214	32.32	87	34	60.92
9	11	− 2	0	0	433	315	27.32	52	28	46.15
7	12	− 1	1	1	431	302	29.85	81	55	32.10
5	13	− 1	− 1	1	328	225	31.32	67	30	55.22
4	14	1	1	− 1	425	238	43.97	72	21	70.83
15	15	0	0	0	429	267	37.81	80	28	65.00
18	16	0	0	0	415	260	37.32	75	28	62.67
12	17	0	2	0	435	306	29.72	71	46	35.21
11	18	0	− 2	0	437	317	27.44	51	30	41.18
2	19	1	− 1	− 1	423	213	49.64	66	16	75.76
19	20	0	0	0	372	241	35.21	83	29	65.06

V_M magnetizing current, T_R rotational speed of tool, W_X working gap, $\%RSR_C$ percentage reduction in surface roughness after CA-BEMRF and *%RRS* percentage reduction in residual stress after CA-BEMRF.

central composite design (CCD) technique are conducted through CA-BEMRF process by varying the input variables, viz. magnetizing current (V_M), Rotational Speed of Tool (T_R), Working gap (W_X) at five different levels as mentioned in Table 4. Each specimen is submerged in chemical agent for a period of 30 minutes to obtain a passive layer in the surface before the mechanical abrasion through MR fluid. Table 5 represents the result of percentage reduction in surface roughness and residual stress after CA-BEMRF process on Al 7075 workpiece.

5.1 Surface Roughness of Al 7075 After CA-BEMRF

A quadratic model was suggested by Design Expert 6.0.8 software for percentage reduction in surface roughness (%RSR$_C$) after CA-BEMRF having F-value of 20.12 and the corresponding p-value which is less than 0.0001. After excluding the non-significant parameters, an empirical relationship (Eq (6)) between the output response and the input process parameters after CA-BEMRF is acquired. With the prediction of the ideal values for the input process parameters, this equation may be used to achieve the maximum percentage reduction in surface roughness when CA-BEMRF of Al 7075 alloy.

$$\begin{aligned} \mathrm{RSR}_C = & 36.264 + 2.989 * V_M + 0.021 * T_R \\ & - 6.441 * W_X - 1.548 * T_R^2 + 1.887 * W_X^2 \\ & - 3.749 * V_M * T_R - 2.699 * T_R * W_X \end{aligned} \quad \text{(Eq 6)}$$

Figure 5(a) interaction graph between V_M and T_R on %RSR$_C$ demonstrates unmistakably that when V_M is maintained at the greatest levels of 3.5A, while T_R is kept at 500 rpm and W_X is maintained at 1.5 mm, %RSR$_C$ achieves a maximum value of 45.83%. This is because hydrogen peroxide used as a chemical agent which softens the surface of the aluminum workpiece and high finishing force is exerted due to increased magnetizing current resulting in better surface finish.

The combined effect of T_R and W_X on %RSR$_C$ as represented in Fig. 5(b) shows that %RSR$_C$ value reduced from 44.15 to 34.4% on increasing T_R from 300 to 700 rpm while keeping W_X at lowest level of 0.5 mm. It is also observed that %RSR$_C$ value decreased to 33.3% from 44.15% once the working gap is increased to the highest value of 2.5 mm, keeping rotational speed of tool constant at 300 rpm.

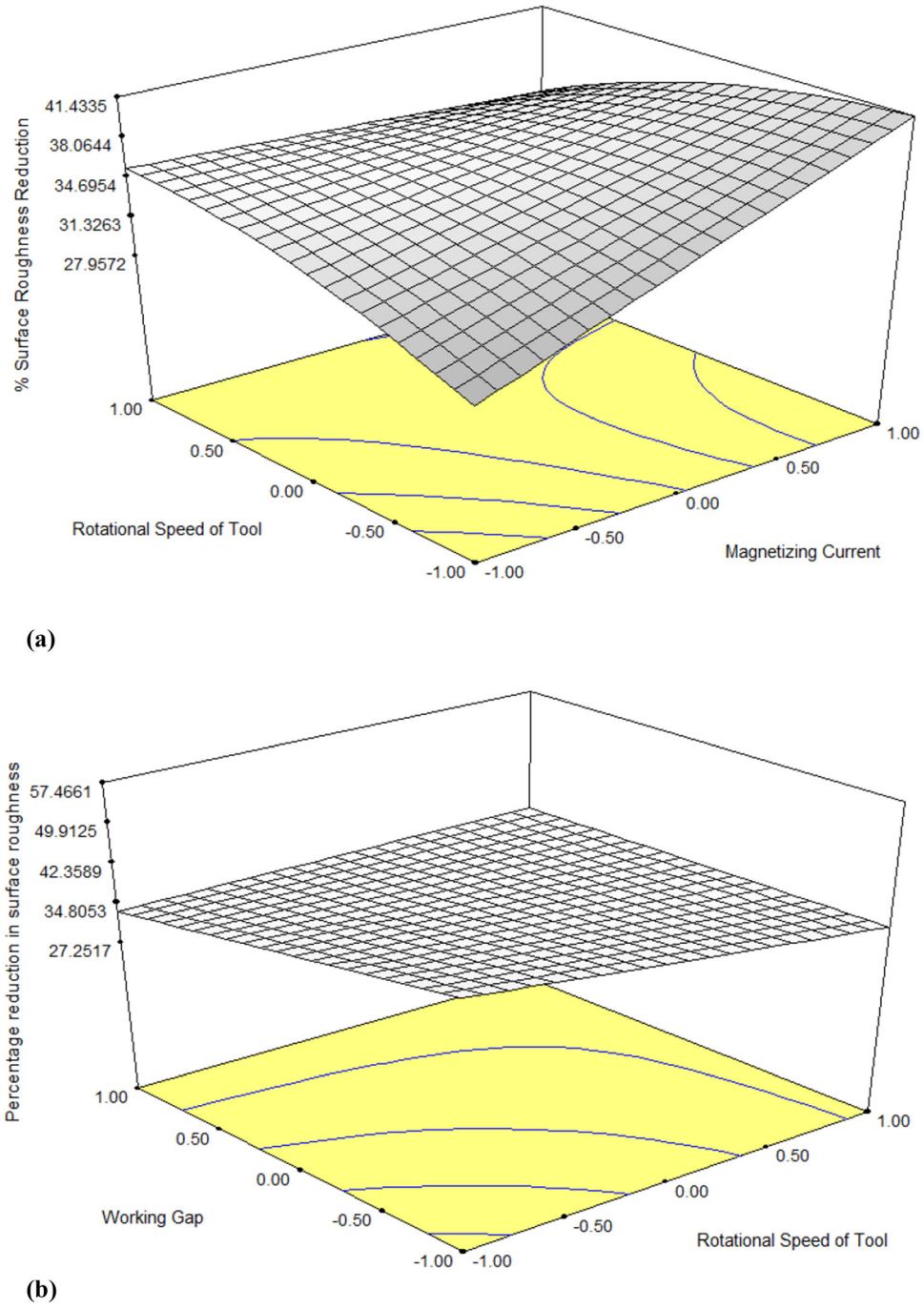
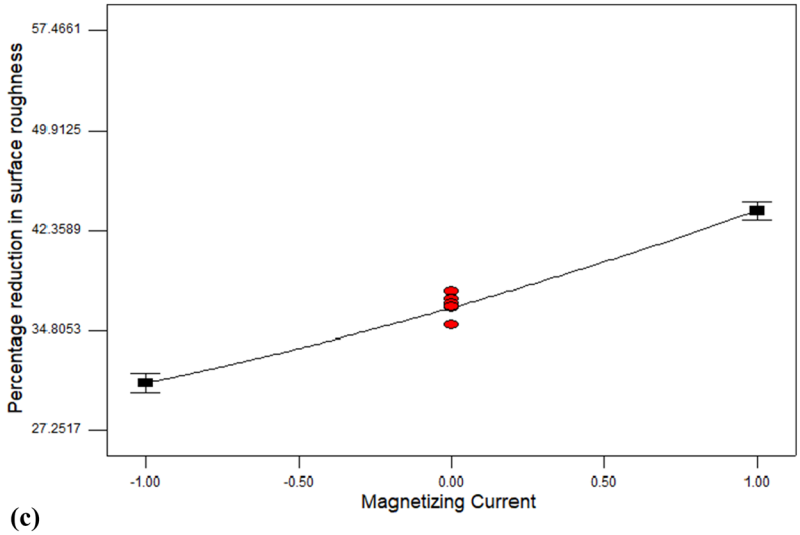


Fig. 5 (a) V_M & T_R interaction graph on percent %RSR$_C$, (b) Interaction graph of T_R & W_X on %RSR$_C$ (c) Individual effect of V_M on %RSR$_C$, (d) Individual effect of T_R on %RSR$_C$ and (e) Individual effect of W_X on %RSR$_C$

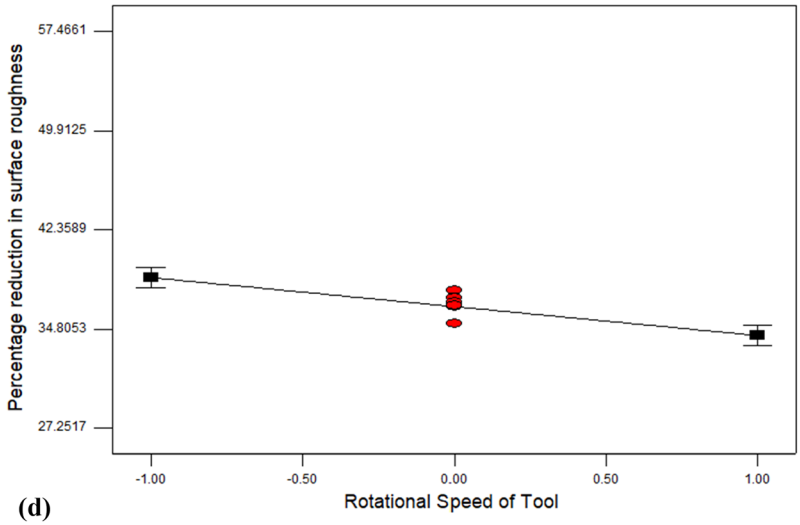
The individual effects of chemical concentration, magnetizing current, tool rotation speed, and working gap on %RSR$_C$ are each depicted in Fig. 5(c), (d) and (e), respectively. The individual effect plot had the same slant as the interaction graph and demonstrated that an increase in V_M raised the %RSR$_C$ while an increase in T_R and W_X decreased the %RSR$_C$ for the previously mentioned reasons. The graphs show that the conditions of 5w/w% chemical composition, 3.5A magnetizing current, 500 rpm rotational speed of the tool, and 1.5 mm working gap result in the highest percentage reduction in surface roughness value of 43.9%.

5.2 RESIDUAL Stress of Al 7075 After CA-BEMRF

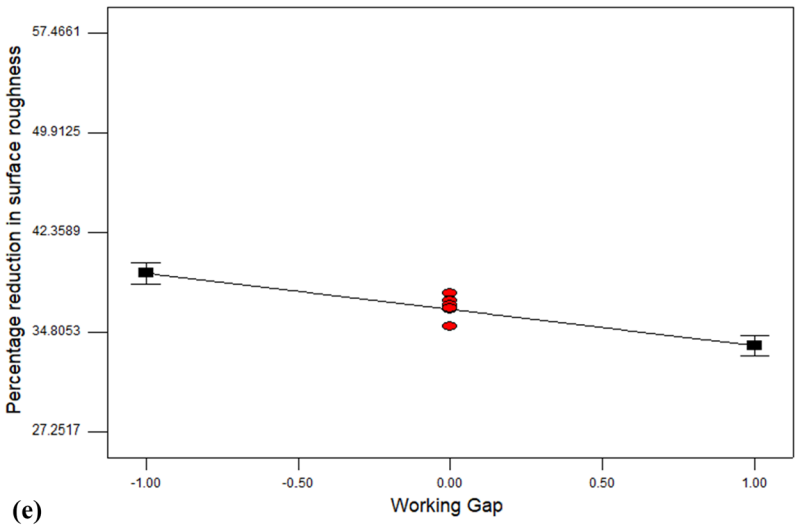
The results of the percentage reduction in residual stress of Al 7075 alloy workpiece after finishing through CA-BEMRF is shown in Table 6. The quadratic model is significant, as evidenced by its F-value of 19.958 and matching p-value of less than 0.001. The anode used in our system is the workpiece. We carefully selected this anode to ensure optimal performance and accurate measurement outcomes. The computer, sensor unit, and power unit are the major three components of the machine. The system comprises of a 30-KV x-ray tube, a 2-D x-ray sensor, and a power supply unit to generate x-rays. One



(c)



(d)



(e)

Fig. 5 continued

Table 6 10 best solutions for machining of Al 7075 using CA-BEMRF

Serial no.	V_M	T_R	W_X	%SRR$_C$	%RRS	Desirability	
1	2.3	550	0.5	58.728	80.542	1	Selected
2	3.3	510	0.6	57.732	88.457	1	Selected
3	3.4	450	0.6	59.589	83.124	1	
4	3.3	495	0.7	59.973	89.183	1	
5	3.4	470	0.8	60.717	87.568	1	
6	3.3	500	0.7	59.677	92.001	1	
7	2.6	540	0.5	59.028	87.765	1	
8	3.1	530	0.6	59.079	89.477	1	
9	3.4	400	2.5	55.138	67.531	0.844	
10	3.5	450	1.6	49.642	66.673	0.747	

reading of residual stress is taken by an x ray in 90 seconds. The machine features an air conditioning element that makes it very effective. The sample was 38 mm away from the x-ray focusing lens, and the X 93 ray tube current was 1 mA. The specimen to be studied was given information on its lattice constant, inter-planar spacing, and diffraction planes (h, k, and l), as well as its crystal structure, Poisson's ratio, and Young's modulus. The incidence angle for x-rays was fixed at 35 degrees. The x-ray machine required an AC supply with the specifications of 240 V and 50/60 Hz to function, although it could alternatively be powered by a battery-powered 24 V source. A unique air-cooling element that was included with the machine ensured its efficient operation. Between the sample and the machine's focusing lens, the incidence angle for x-rays striking the sample surface was adjusted at 35°. Lattice constant 2.86A; inter-planar distance 1.17 m; diffraction plane 2, 1, 1; diffraction angle156.390; crystal structure BCC; Poisson's ratio 0.28; the x-axis Cosα represents cos of azimuth angle of Debye sheerer ring and y-axis represents strain ($\varepsilon\alpha$1). The inclination angle of bend was measured as the residual stress. This selection is integral to achieving meaningful and accurate results in our experiments.

After removing the non-significant process factors, an empirical relationship (Eq (7)) is found between the output response and the input process parameters. By forecasting the ideal values of the input process parameters, this equation can be used to achieve the maximum decrease in residual stress during BEMRF of Al7075 material.

$$\%\text{RRS} = 63.782 + 4.418 * V_M - 3.926 * T_R - 9.727 * W_X - 6.079 * C^2 - 7.519 * T_R * W_X \quad \text{(Eq 7)}$$

After the initial grinding process a high level of residual stress of 46 MPa as induced in the workpiece surface as shown in Fig. 6. The residual stresses on the aluminum 7075 workpiece obtained after the CA-BEMRF process

are shown in Fig. 7 while the tool's rotating speed, working gap, and magnetizing current are maintained at 2.5 mm, 500 rpm, and 0.5 mm, respectively. Although there is a grain size effect in the residual stress before and after CA-BEMRF process, with the aforementioned parametric parameters, a residual stress reduction of up to 78.26% is achieved.

The combined effect of T_R and W_X on %RRS is shown in Fig. 8(a), and it demonstrates that when T_R is increased from 300 to 700 rpm while W_X is maintained at 2.5 mm, the value of %RRS decreases from 59.42 to 36.53%. Contrarily, when W_X was reduced from 2.5 to 0.5 mm and T_R remained constant at 700 rpm, the percentage of residual stress that was reduced dramatically increased to a high value of 71.02%. It was shown that the workpiece experiences shear and Hertzian stress as a result of vibrations with the abrasive particles when the tool's rotating speed increases while the working gap decreases.

The separate effects of V_M, T_R, and W_X on %RRS are depicted in Fig. 8(b), (c), and (d). As seen in Fig. 8(b), an increase in magnetizing current causes a greater percentage reduction in residual stress. The results show that %RRS increases from 59.36 to 68.2% on increasing V_M to 3.5A from 0.5A while keeping T_R and W_X constant at 500 rpm and 1.5 mm, respectively. The individual effect plot, which had the same slant as the interaction graph of T_R and W_X, indicated that increasing T_R and W_X decreases the %RRS.

The plastic deformation occurs at the point of maximum shear stress and thus, since layers below the surface elongate more, the maximum residual stress is observed below the surface. In this case, however, other parameters which are proportional to plastic deformation would be expected to exhibit a maximum at the same depth as the residual stresses.

An alternative explanation is as follows. If the layers near the surface slip with respect to each other because of the high intensity of the operation, i.e., there is unconstrained flow which decreases with depth, the residual stresses will exhibit a maximum value below the surface. With the maximum total plastic deformation occurring at the surface. The depth at which the residual stress is maximum is the depth at which maximum constrained plastic flow occurs.

6. Output Response Optimization

A desirability function approach was presented by Derringer (Ref 25) and used to optimize the response to the output i.e., percentage surface roughness reduction and percentage reduction in residual stress after CA-BEMRF (%SRR$_C$). This method assigns a dimensionless desirability value to the measured features of each expected response where d varies between 0 and 1. According to Montgomery (Ref 26), if the desirability function reaches a value of 0, it denotes an undesired response, and if it reaches a value of 1, the response is exactly what was intended. The value of d increases in direct proportion to how desirable the associated response is. The greater the scenario examined in the current work, the better the percentage reduction in surface roughness. The response is converted into a dimensionless function di using Eq (8) shown as follows:

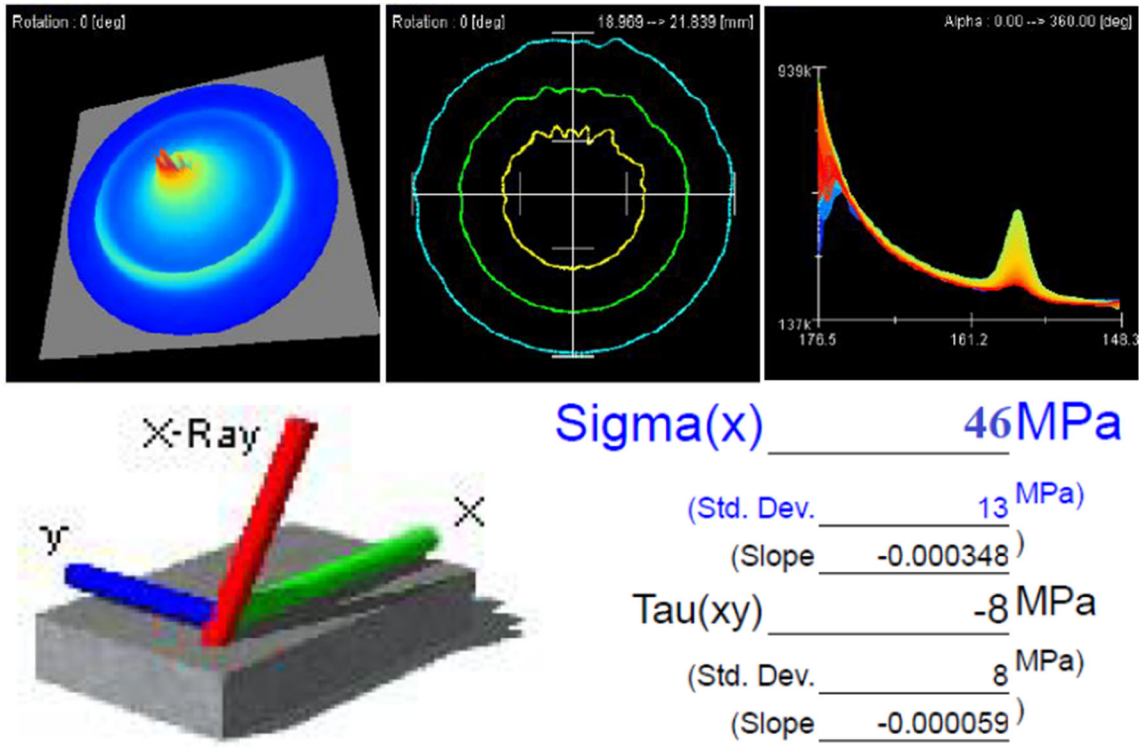


Fig. 6 Residual stress on aluminum 7075 workpiece after grinding process

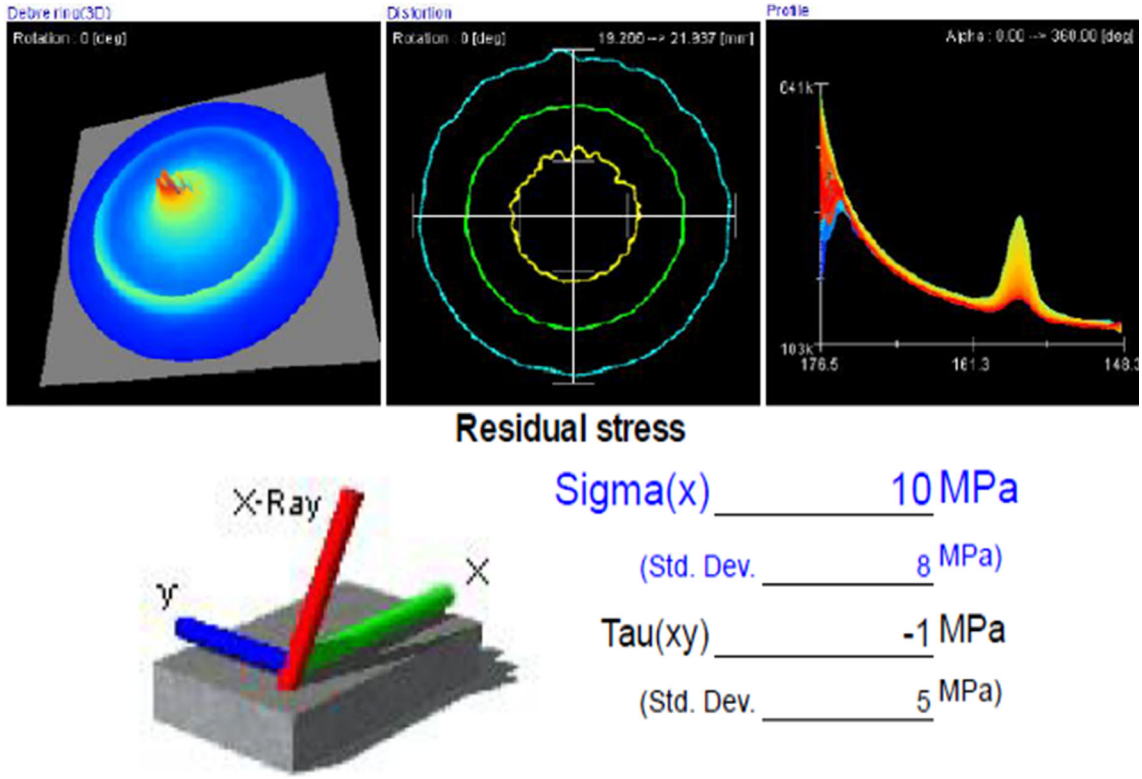


Fig. 7 Residual stress on aluminum 7075 workpiece after CA-BEMRF process

$$d_i = \begin{cases} 0 & Xi < Bi \\ (Xi - Bi/Hi - Bi)^{w} & Bi \le Mi \le Hi \\ 1 & Bi > Hi \end{cases} \quad \text{(Eq 8)}$$

where Bi = Lower response limit, Mi = Upper response limit, Hi = Response goal value. B and M are selected in RSM according to the mathematical.

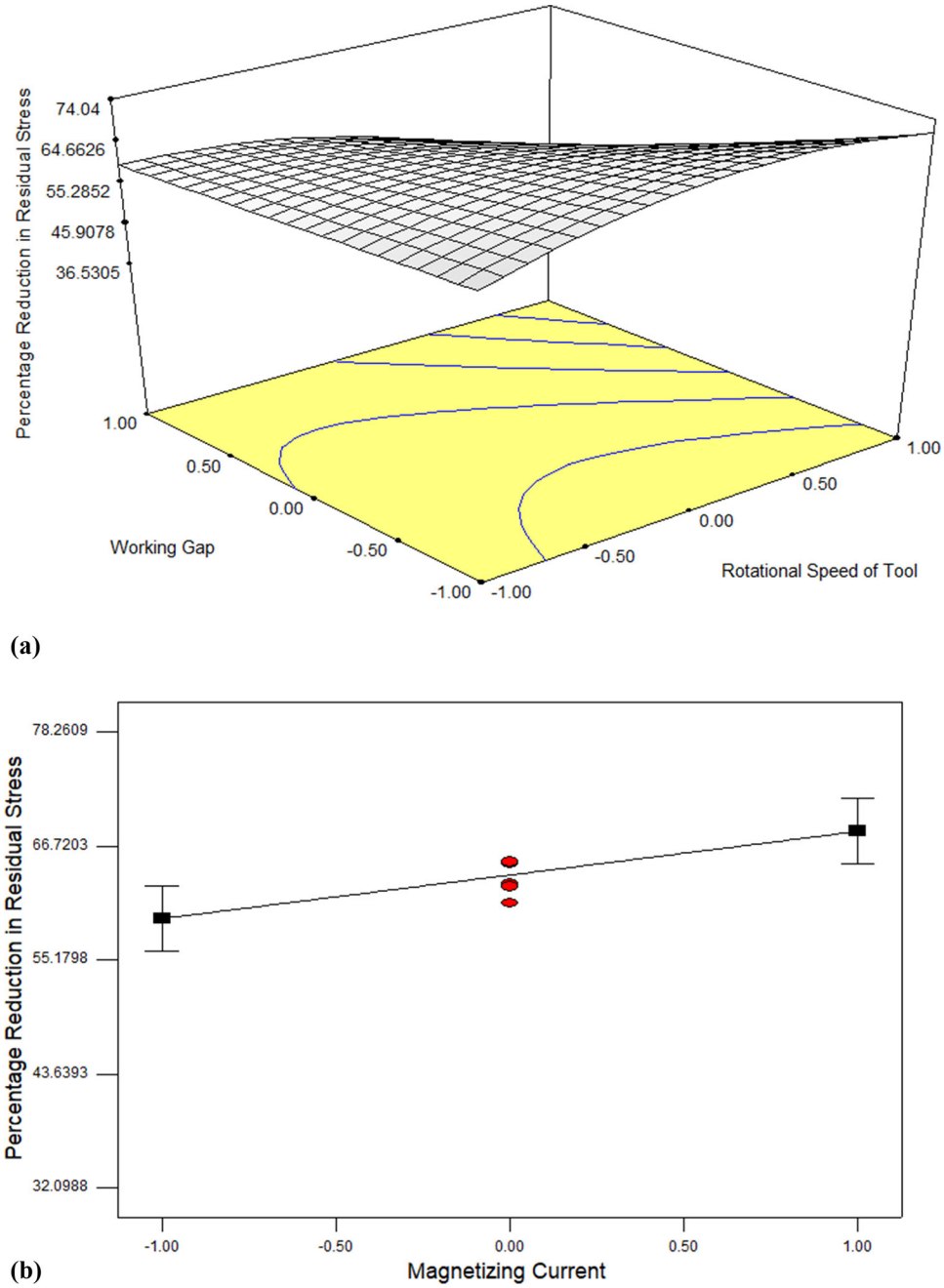


Fig. 8 (a) Interaction graph of T_R and W_X on %RRS, (b) Individual effect of V_M on %RRS, (c) Individual effect of T_R on %RRS and (d) Individual effect of W_X on %RRS

Determining the desirability response value and optimizing the desirability function to find the best potential response value are the first two steps in optimization.

Design Expert 6.0.8 software has been used to perform response optimization. Table 6 shows the 10 best solutions obtained using the desirability method. The highest desirability value obtained is 1, which means that when finishing Al 7075, the maximum percentage reduction in surface roughness and residual stress through CA-BEMRF process can be obtained at V_M = 2.3A, T_R = 550 rpm and W_X = 0.5 mm.

A confirmation experiment was performed for solutions no. 1 and 2 to verify the above solutions. Table 7 summary of the findings from the confirmatory experiments shows that the error was less than 1.5%. These findings support the high repeatability of the beneficial optimization-derived solutions.

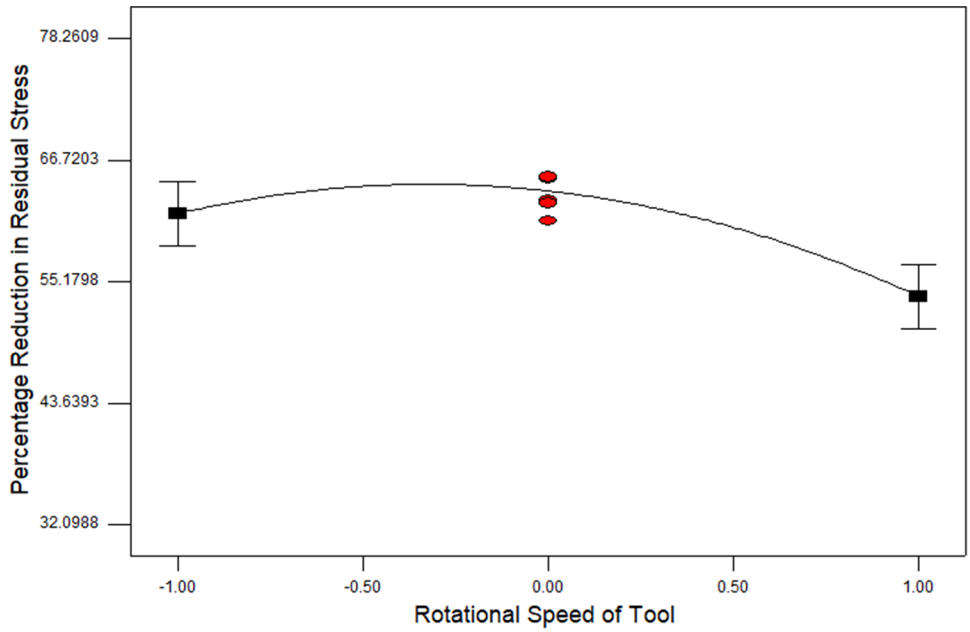
7. Surface Topography of Machined Surface

Scanning electron micrograph (SEM) of aluminum 7075 alloy workpiece surface after grinding process and CA-BEMRF process is shown in Fig. 9(a) and (b), respectively.

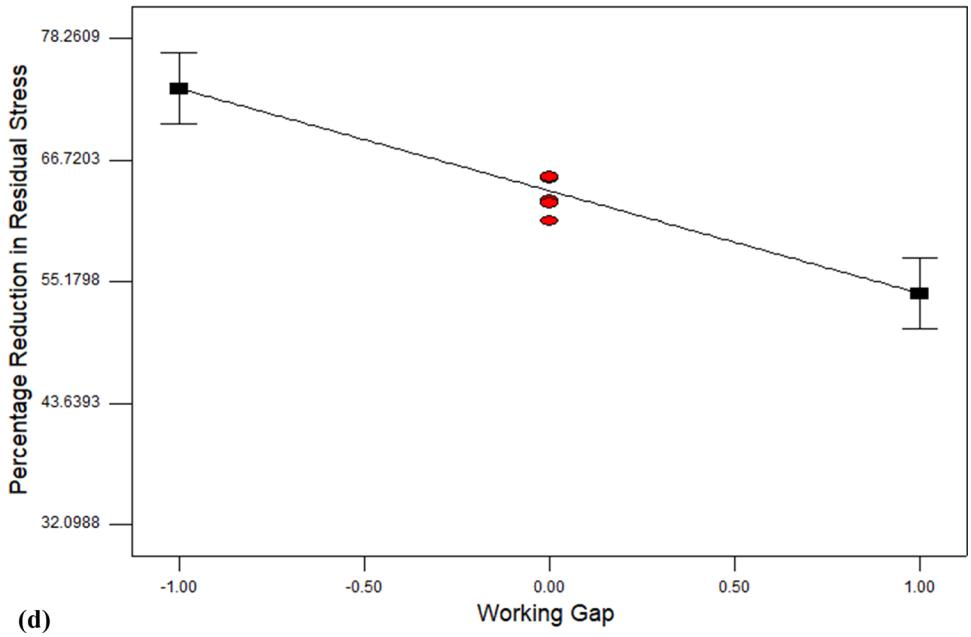
The lays are visible clearly on the workpiece surface due to the pre-finishing technique i.e., grinding process as shown in Fig. 9(a). It is observed from Fig. 9(b) that more uniform and improved surface finish with very fewer grinding marks is

Table 7 Confirmatory experiments

Serial no.	V_M	T_R	W_X	Predicted		Actual		Error %	Error %
				$\%RSR_C$	%RRS	$\%RSR_C$	%RRS	$\%RSR_C$	%RRS
1	2.3	550	0.5	58.728	80.542	59.171	81.781	− 0.75	− 1.24
2	3.3	510	0.6	57.732	88.457	57.827	89.168	− 0.16	− 0.8



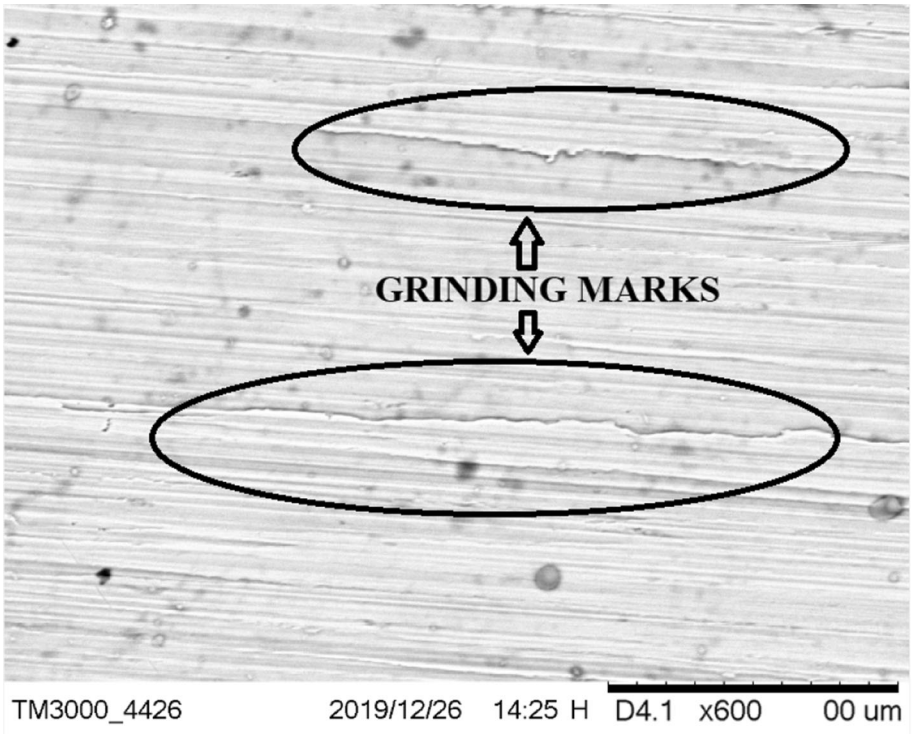
(c)



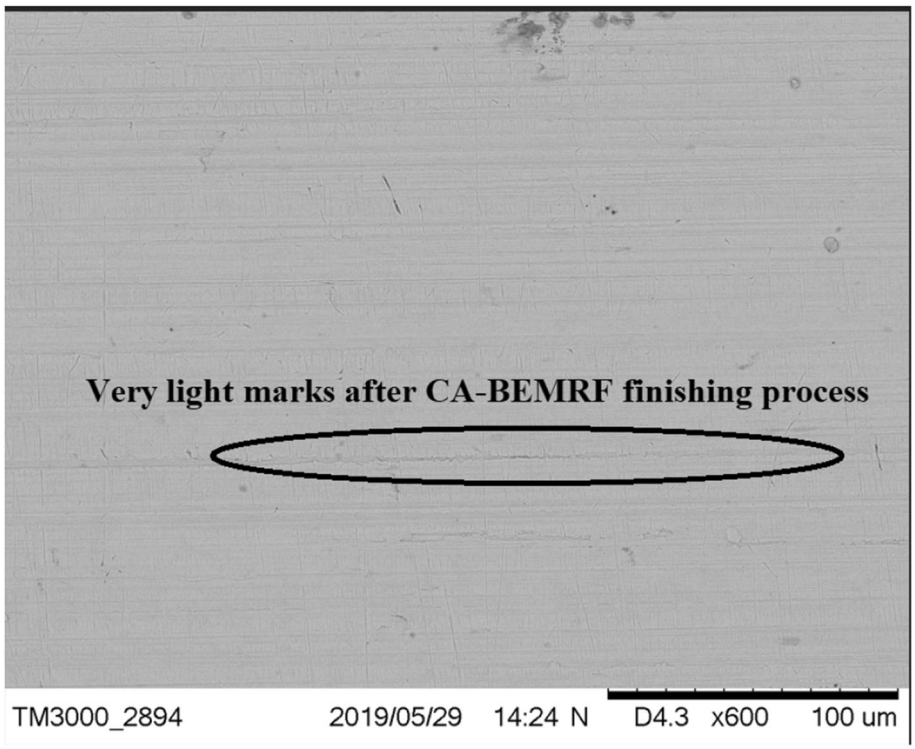
(d)

Fig. 8 continued

obtained after CA-BEMRF of the workpiece. When finished with the assistance of chemical agent hydrogen peroxide, percentage reduction in surface roughness increased to 57.32% leading to an improved surface finish as shown in Fig. 9(b). This finish was obtained at 5w/w% chemical composition, 2.5A magnetizing current, 500 rpm rotational speed of tool and 0.5 mm working gap during CA-BEMRF process.



(a)



(b)

Fig. 9 SEM of the workpiece after (a) grinding process and (b) CA-BEMRF process

A 59.171% reduction in surface roughness and 81.871% reduction in residual stress was obtained after the confirmatory experiment conducted at 5 w/w% chemical composition, 2.3A magnetizing current, 550 rpm rotational speed of tool and 0.5 mm working gap. SEM obtained for this specimen was found better as shown in Fig. 10 when compared with the SEM image obtained after grinding process shown in Fig. 9(a).

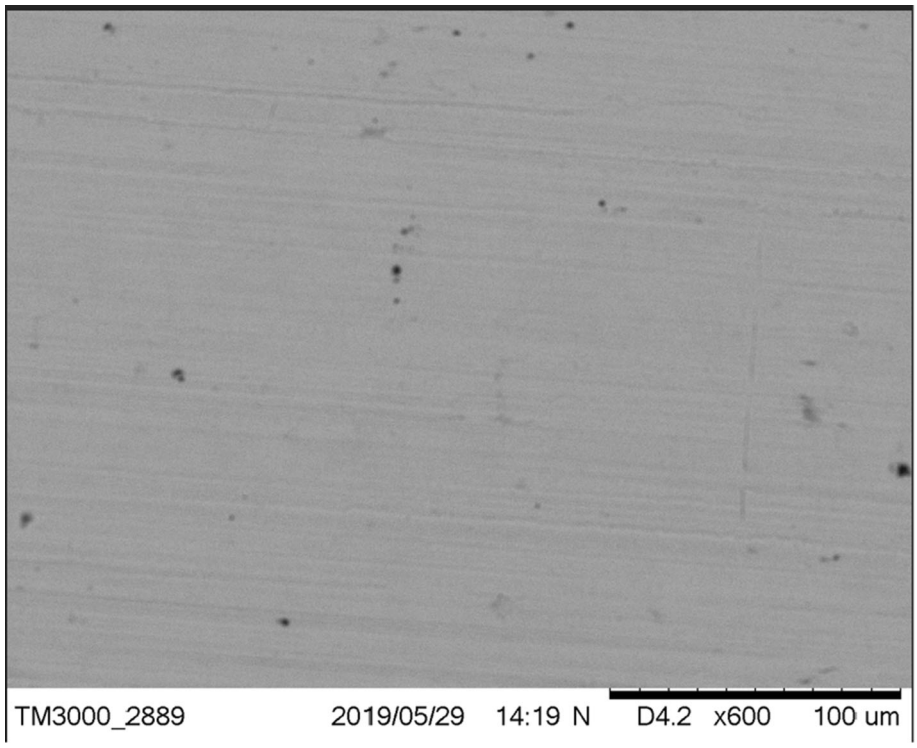


Fig. 10 SEM of workpiece after confirmatory experiments

8. Conclusion

This study investigates how to reduce surface roughness and residual stress on the surface of an Al 7075 workpiece caused by grinding by employing the CA-BEMRF technique, which combines mechanical and chemical processes. The effects of input constraints, such as magnetizing current, tool rotation, and working gap, were studied on percentage surface roughness reduction and percentage reduction in residual stress. Additionally, the impact of chemical composition on the percentage reduction in surface roughness reduction after the CA-BEMRF process ($\%SRR_C$) was investigated. With design expert software, the optimization of the percentage reduction in surface roughness was carried out using the desirability technique. The following conclusions are drawn from this investigation.

- The experiments have been conducted on aluminum 7075 alloy workpiece with CA-BEMRF process.
- The CA-BEMRF process was able to provide a better surface finish and reduced residual stresses.
- The residual stress induced by the grinding process was significantly reduced by using CA-BEMRF process.
- Experimental results showed that the average percentage reduction in surface roughness of 38.19% was achieved when Al 7075 alloy workpiece was finished using CA-BEMRF process.
- The results of the confirmatory experiments conducted were found very close to the predicted outcome with a maximum of − 1.24% error.
- The study also provides a set of optimized parameters for CA-BEMRF process to achieve high level surface finish and minimum residual stress for Al 7075 alloy workpiece.
- The study found that the chemical-assisted BEMRF process produced a significantly smooth and more consistent surface finish on the aluminum alloy.
- SEM images of the finished workpieces showed an improved surface finish after CA-BEMRF process.

References

1. A. Sharma, and M.S. Niranjan, Experimental and theoretical investigation into surface roughness and residual stress using ball end magnetorheological finishing, *J. Eng. Res*., 2022
2. S. Jha, and V.K. Jain, *Micromanufacturing and Nanotechnology*, Springer, Berlin, 2006
3. M.S. Tov, and J.A. Crabtree, Personal cooling apparatus and method, Patent US:6276155B2. 2001
4. S.Y. Chiu, Y.L. Wang, C.P. Liu, J.K. Lan, C. Ay, M.S. Feng, and B.T. Dai, The application of electrochemical metrologies for investigating chemical mechanical polishing of Al with a Ti barrier layer, *Mater. Chem. Phys.*, 2003, **82**(2), p 444–451.
5. P. Ranjan, R. Balasubramaniam, and V.K. Suri, Development of chemo-mechanical magnetorheological finishing process for super finishing of copper alloy, *Int. J. Manuf. Technol. Manage.*, 2013, **27**(4–6), p 130–141.
6. S. Babu, S. Hegde, S. Jha, U. Patri, and Y. Hong, Slurry and method for chemical-mechanical planarization of copper, Patent US:20050026444 A1. 2005
7. M. Boccas, T. Vucina, C. Araya, E. Vera, and C. Ahhee, Protected-silver coatings for the 8-m Gemini telescope mirrors, *Thin Solid Films*, 2006, **502**(1–2), p 275–280.
8. R. Almanza, P. Hernández, I. Martínez, and M. Mazari, Development and mean life of aluminum first-surface mirrors for solar energy applications, *Sol. Energy Mater. Sol. Cells*, 2009, **93**(9), p 1647–1651.
9. J. Casstevens, *A 3 Meter Capacity Diamond TURNING machine will be Enabling Technology for Low Cost Fabrication of x-ray Telescope Optical Components*, Optical Systems Inc, Dallas, 2011

10. A. Gebhardt, J. Kinast, R.R. Rohloff, W. Seifert, M. Beier, S. Scheiding, and T. Peschel, Athermal metal optics made of nickel plated AlSi40. In *International Conference on Space Optics—ICSO,* 2017, vol. 10563, pp. 451–459
11. A. Rathod, V. Ghai, P. Parmar, S. Mishra, S. Ghildiyal, P. Ranjan, and R. Balasubramaniam, Development of intrinsic fabry perot interferometic sensor based on a single mode–multi mode–single mode (SMS) configuration. *COPEN-8*., 2013
12. Y. Ahn, J.Y. Yoon, C.W. Baek, and Y.K. Kim, Chemical mechanical polishing by colloidal silica-based slurry for micro-scratch reduction, *Wear*, 2004, **257**(7–8), p 785–789.
13. M. Shen, F. Xiao, and K. Alameh, A novel reconfigurable optical interconnect architecture using an Opto-VLSI processor and a 4-f imaging system, *Opt. Express*, 2009, **17**(25), p 22680–22688.
14. V.K. Jain, P. Ranjan, V.K. Suri, and R. Komanduri, Chemo-mechanical magneto-rheological finishing (CMMRF) of silicon for microelectronics applications, *CIRP Ann.*, 2010, **59**(1), p 323–328.
15. P. Ranjan, R. Balasubramaniam, and V.K. Suri, Modelling and simulation of chemo–mechanical magnetorheological finishing (CMMRF) process, *Int. J. Precis. Technol.*, 2014, **4**(3–4), p 230–246.
16. P. Ranjan, R. Balasubramaniam, and V.K. Jain, Analysis of magnetorheological fluid behavior in chemo-mechanical magnetorheological finishing (CMMRF) process, *Precis. Eng.*, 2017, **49**, p 122–135.
17. V. Ghai, P. Ranjan, A. Batish, and H. Singh, Atomic-level finishing of aluminum alloy by chemo-mechanical magneto-rheological finishing (CMMRF) for optical applications, *J. Manuf. Process.*, 2018, **32**, p 635–643.
18. Y. Kumar, and H. Singh, Chemomechanical magnetorheological finishing: process mechanism, research trends, challenges and opportunities in surface finishing, *J. Micromanuf.*, 2022, **5**(2), p 193–206.
19. H.S. Kuo, and W.T. Tsai, Effects of alumina and hydrogen peroxide on the chemical-mechanical polishing of aluminum in phosphoric acid base slurry, *Mater. Chem. Phys.*, 2001, **69**(1–3), p 53–61.
20. A.K. Singh, S. Jha, and P.M. Pandey, Magnetorheological ball end finishing process, *Mater. Manuf. Processes*, 2012, **27**(4), p 389–394.
21. C. Qian, Z. Fan, Y. Tian, Y. Liu, J. Han, and J. Wang, A review on magnetic abrasive finishing, *Int. J. Adv. Manuf. Technol.*, 2021, **112**, p 619–634.
22. C. Qian, Y. Tian, Z. Fan, Z. Sun, and Z. Ma, Investigation on rheological characteristics of magnetorheological shear thickening fluids mixed with micro CBN abrasive particles, *Smart Mater. Struct.*, 2022, **31**(9), p 095004.
23. Z. Sun, Z. Fan, Y. Tian, C. Prakash, J. Guo, and L. Li, Post-processing of additively manufactured microstructures using alternating-magnetic field-assisted finishing, *J. Market. Res.*, 2022, **19**, p 1922–1933.
24. Z. Fan, Y. Tian, Q. Zhou, and C. Shi, Enhanced magnetic abrasive finishing of Ti–6Al–4V using shear thickening fluids additives, *Precis. Eng.*, 2020, **64**, p 300–306.
25. G. Derringer, and R. Suich, Simultaneous optimization of several response variables, *J. Qual. Technol.*, 1980, **12**(4), p 214–219.
26. D.C. Montgomery, *Design and Analysis of Experiments*, Wiley, New York, 2001

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Physica Scripta

INCLUSIVE PUBLISHING
TRUSTED SCIENCE

PAPER

Dy^{3+} doped $KCa(PO_3)_3$ phosphor for white light generation: structural and luminescent studies

To cite this article: Deepti *et al* 2024 *Phys. Scr.* **99** 065573

View the article online for updates and enhancements.

You may also like

- NMR and NQR studies on transition-metal arsenide superconductors $LaRu_2As_2$, $KCa_2Fe_4As_4F_2$, and $A_2Cr_3As_3$
Jun Luo, , Chunguang Wang et al.
- ^{19}F NMR Study of the Bilayer Iron-Based Superconductor $KCa_2Fe_4As_4F_2$
Yu-Ting Shao, , Wen-Shan Hong et al.
- Distinct evolution of charge occupations on Fe 3d orbitals in hole-doped iron arsenic superconductors
Ming-Cui Ding, Li-Min Wu, Gang Zhao et al.

This content was downloaded from IP address 14.139.251.98 on 11/06/2024 at 11:23

Physica Scripta

PAPER

Dy^{3+} doped $KCa(PO_3)_3$ phosphor for white light generation: structural and luminescent studies

RECEIVED
29 February 2024

REVISED
26 April 2024

ACCEPTED FOR PUBLICATION
20 May 2024

PUBLISHED
30 May 2024

Deepti[1], **Sandip Maurya**[1], **Sheetal Kumari**[1], **Pooja Rohilla**[1], **Aman Prasad**[2] **and A S Rao**[1,*]

[1] Department of Applied Physics, Delhi Technological University, Bawana Road, Delhi, 110042, India
[2] Department of Physics and Computer Science, Dayalbagh Educational Institute (DEI), Deemed University, Dayalbagh, Agra, 282005, India
[*] Author to whom any correspondence should be addressed.

E-mail: drsrallam@gmail.com

Keywords: XRD, phosphor, photoluminescence, CIE coordinates

Abstract

Using the traditional solid-state reaction approach, Dy^{3+} ions doped $KCa(PO_3)_3$ phosphors have been synthesized to investigate their luminescent properties to produce high-quality white light for solid-state lighting applications, particularly in white LEDs. x-ray diffraction (XRD) patterns were used to examine the structure and phase of the phosphors. Using scanning electron microscopy (SEM), the morphology of the as-synthesized phosphor has been investigated. Fourier transform infrared spectroscopy (FT-IR) has been used to investigate several vibrational bands seen in the phosphor. Using diffuse reflectance spectra (DRS), the as-synthesized phosphors' optical band gap values have been estimated. When Dy^{3+} ions are excited at 350 nm, the photoluminescence (PL) spectra characteristics observed for the activated $KCa(PO_3)_3$ phosphor show strong white area emission due to both 575 nm, which is related to the $^4F_{9/2} \rightarrow {}^6H_{13/2}$ and 482 nm that is ascribed to $^4F_{9/2} \rightarrow {}^6H_{15/2}$ transition of Dy^{3+} ions. Additionally, the concentration quenching of Dy^{3+} ions doped at 4 mol% is seen in the PL spectra. Based on the observed PL spectra, the computed CIE chromaticity coordinates for the optimised $KCa(PO_3)_3$ phosphors are located in the deep white area. The lifespan of the as-titled phosphors decreases as the amount of Dy^{3+} ions increases in the host lattice. Additionally, the PL decay profiles shows a dual exponential behaviour when excited at 350 nm, with an emission wavelength at 575 nm. The lifetime values were used to calculate the quantum efficiency of the as prepared phosphors. On the basis of the results of the aforementioned studies, we wish to project Dy^{3+} ion doped $KCa(PO_3)_3$ phosphors as white light generating materials in w-LEDs and for other photonic applications.

1. Introduction

Phosphors are widely used in solid-state lighting, display devices, and anti-counterfeiting applications [1, 2]. White light-emitting diodes (w- LEDs) have gained popularity in recent years due to their low power consumption and environmental benefits. They are expected to replace traditional light sources such as tungsten halogen lamps and incandescent lamps. LEDs have a significantly longer lifespan when compared to other lighting technologies like Low-Pressure Sodium (LPS) and fluorescent lights. The longevity of LEDs far surpasses that of incandescent lights. The latest LED lights can last for over 50,000 to 100,000 h or more. In contrast, incandescent bulbs have a lifespan that is only 1%–5% as long at best, lasting for approximately 1,200 h. The additional features of these properties include a broad spectrum of colors that span from ultraviolet to infrared, a limited spectral bandwidth, a high degree of luminous efficiency, and an extremely rapid response time [3, 4]. The most common method for creating white LEDs is to combine phosphors with UV or n-UV chips made of InGaN. The development process of the standard LED involves three crucial steps: seasoning, selection, and certification. These steps are meticulously carried out to ensure the optimal performance of the w-LED [5–7]. During seasoning, the LED is subjected to a series of tests to stabilize its electrical and thermal properties. The

© 2024 IOP Publishing Ltd

selection process involves choosing the most suitable w-LED based on its characteristics and performance. Finally, the certified LED undergoes rigorous testing to ensure its compliance with industry standards and specifications. Generally, phosphors consist of an activator, a host material, and sometimes a sensitizer. Rare earth or metals are often used as activator ions for doping. Red, green, blue, and white light can be produced by doping one or more rare earth (RE) ions in an appropriate single matrix. Selecting a suitable substrate material is crucial for obtaining energy-efficient phosphors [1].

Phosphates are crucial luminescent materials that offer various crystal-field environments for emission centers. Phosphates doped with rare-earth-elements (REEs) are thermally stable and have a rigid tetrahedral matrix that is ideal for charge stabilization. Moreover, they are excellent host materials for producing phosphors as they require low sintering temperatures [8]. Researchers have recently explored alkali and alkaline-earth-metal phosphate phosphors to develop new materials for UV-based w-LEDs. Many studies have focused on $AB(PO_3)_3$ phosphates, with A representing monovalent cations (Li^+, Na^+, K^+, and Cs^+) and B representing divalent cations (Ca^{2+}, Sr^{2+}, and Ba^{2+}) [9, 10]. These phosphates have efficient thermal and hydrolytic stabilities, making them ideal host materials for w-LEDs.

$KCa(PO_3)_3$ has a unique structure comprising of twisted polyphosphate rings forming a cyclic anion $(P_{12}O_{36})^{12-}$, where non-bridging oxygen atoms point toward the center [11, 12]. This structure creates a distorted octahedral coordination environment for the central potassium ions and distorted octahedral cavities for calcium ions between the rings. $KCa(PO_3)_3$ possesses desirable properties such as low phonon energy, high optical damage threshold, good charge stabilization, and chemical stability, making it a suitable host material for lanthanide activators [13]. Additionally, $KCa(PO_3)_3$ exhibits transparency in the spectra range of interest and low sintering temperature. Ca-based phosphate compounds are of interest to researchers due to their significant UV absorption and serving as an important host for luminescence [14].

Dysprosium (Dy^{3+}) doped luminescence materials have been gaining increasing attention for their superior white emissions [15, 16]. When incorporated in various host matrices, Dy^{3+} ion is known to display two prominent emission bands in its spectra. The first band corresponds to the pseudo-quadrupole transition $^4F_{9/2} \rightarrow {}^6H_{13/2}$, which involves a change in both ($\Delta L = 2$) and ($\Delta J = 2$). This transition occurs at a wavelength of 576 nm and is characterized by a yellow emission color, while the second band is associated with the transition $^4F_{9/2} \rightarrow {}^6H_{15/2}$ at 478 nm (blue) [17, 18]. One can achieve almost white emission using Dy^{3+} doped luminescent materials by manipulating the yellow-bluish intensity ratio (Y/B). Therefore, Dy^{3+-} -activated phosphors have emerged as a promising option for developing white light [2, 19]. Researchers have found that Dy^{3+}-activated phosphors offer advantages over other options in terms of their ability to produce white light. Therefore, they are considered a strong candidate for use in a variety of applications.

In our work, we have synthesized a series of polyphosphates, or $KCa(PO_3)_3$, (KCP) doped with Dy^{3+} phosphors. This study examines the structural and luminescent properties of KCP: Dy^{3+} phosphors, emphasizing the analysis of the concentration quenching mechanism, luminescence decay characteristics, and color perception. The ultimate goal of this study was to achieve white light emission through the utilization of the prepared phosphors. The results of this examination hold great potential for making significant advancements in the field of materials science.

2. Experiment section

2.1. Sample preparation

A series of Dy^{3+} doped phosphors $KCa(PO_3)_3$ at different concentrations of Dy^{3+} ions (i.e. 1.0, 2.0, 3.0, 4.0, and 5.0 mol%) was prepared via solid-state reaction technique using K_2CO_3(99%, Fisher Scientific), $CaCO_3$(98.0%–100.5%, Fisher Scientific), $NH_4H_2PO_4$(99%, Fisher Scientific), and Dy_2O_3(99.98%, Thermo Scientific) as raw materials. After mixing the raw materials in a stoichiometric ratio with acetone for 1 hour in an agate mortar, the powder was then transferred into the alumina crucible. The resulting powder was heat-treated at 673 K for 3 hours in the furnace to eliminate carbonates. And then, the powder was sintered at 873 K for 5 hours in an air atmosphere. Finally, the powder was ground again in an agate mortar into a fine powder for further characterization.

2.2. Experimental techniques

Using the Bruker Model D8 Advance equipment, which was built to produce nickel-filtered Cu-Kα radiation (0.154 nm), the XRD spectra of the prepared materials was observed in the 2θ range of 20° to 80°. Using a JEOL 7610 F Plus SEM machine, surface morphology was recorded. An un-doped phosphor's FT-IR spectrum was captured with a Perkin-Elmer Frontier FT-IR spectrometer, ranging from 400 to 4000 cm^{-1}. DRS was obtained for the Dy^{3+}-doped phosphors in the 200–1400 nm wavelength range using a JASCO model V-923 UV–Vis–NIR spectrophotometer. A JASCO FP-8300 spectrofluorometer with a resolution of 0.1 nm was used to capture

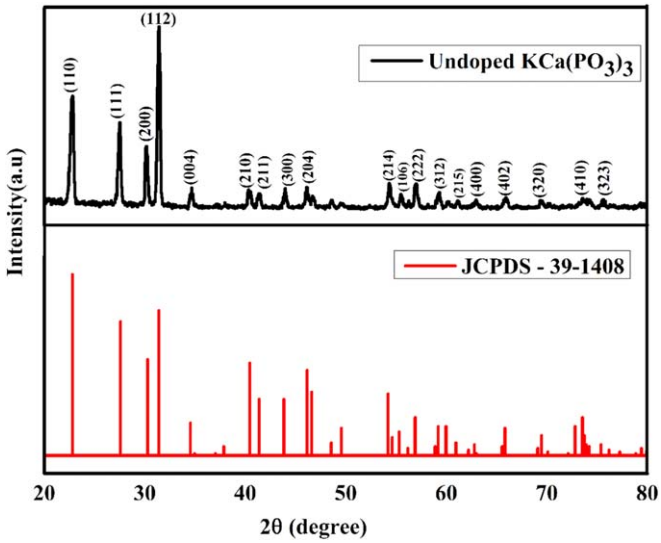


Figure 1. XRD pattern of un-doped $KCa(PO_3)_3$ phosphor along with standard diffraction pattern (JCPDS) 39–1408.

the PL excitation and emission spectra for each Dy^{3+}-ion doped phosphors under a Xenon excitation source. Luminescent decay curves were measured using an Edinburgh FL920 Fluorescence Spectrometer. Each measurement was carried out at room temperature.

3. Results and discussions

3.1. X-ray diffraction (XRD) analysis

The crystallographic arrangement and phase purity of the un-doped $KCa(PO_3)_3$ phosphors were examined using XRD patterns. The hexagonal crystal structure $KCa(PO_3)_3$ (JCPDS No. 39–1408) is a member of the P-6c2 (188) space group. Every diffraction peak is precisely matched to the JCPDS 39–1408 standard card's peaks as shown in figure 1. With the lattice coordinates a = 6.8077 Å, b = 6.8077 Å, and c = 10.3730 Å as well as $\alpha = 90°$, $\beta = 90°$, and $\gamma = 120°$, it suggests the samples are in single-phase. The XRD patterns of the phosphor $KCa(PO_3)_3$ that were in line with the previously described JCPDS database showed that the phosphors had been successfully synthesized with a high level of purity. The addition of Dy^{3+} ions, demonstrates that the synthesized phosphors are devoid of any contaminating phases. One K site and one Ca site were present in the $KCa(PO_3)_3$ host and they both shared the six-fold coordination [20, 21]. Due to the similarity in their ionic radii Ca^{2+}(radii = 1.00 Å) and Dy^{3+}(radii = 0.908 Å), the Dy^{3+} ions chose to occupy the Ca^{2+} site when incorporated into the host, $KCa(PO_3)_3$. The as-titled phosphors exhibit strong and distinct diffraction peaks, indicating their high crystallinity—a prerequisite for comparatively superior luminosity.

3.2. Morphological analysis from SEM

The morphology and particle size information of the prepared phosphors is better understood by examining FE-SEM pictures of an undoped and 4.0 mol% doped $KCa(PO_3)_3$ phosphor. $KCa(PO_3)_3$ powder phosphors have an uneven polyhedron shape and unequal particle size [22], as shown in figures 2(a), (b). The high-temperature reaction method and manual grinding have also resulted in very low agglomeration of some phosphor particles. Most of the phosphor particles available in the market today are between 2 and 10 μm in size. The ImageJ Software is used to find the size of the particle for an un-doped and 4.0 mol% of Dy^{3+} ions doped $KCa(PO_3)_3$ phosphor. The sizes fall in the region of 3–5 μm [18]. As a result ofthe samples' particle sizes being in the micrometer range, these materials can be aptly used for coating white LEDs and display devices [23]. Furthermore, energy dispersive x-ray analysis (EDX) has been used to analyze the elemental makeup of the as-synthesized phosphors. The EDX for the doped $KCa(PO_3)_3$: 4 mol% Dy^{3+} doped phosphor is presented in figure 3. Every precursor elements is visible in the phosphor lattice, according to the EDX picture.

3.3. FT-IR analysis

Vibrational spectroscopy (FT-IR) was conducted to investigate the bonding properties of un-doped $KCa(PO_3)_3$ phosphor. The vibrational spectrum of the prepared phosphor shown in figure 4 was recorded from 400 cm^{-1} to 4000 cm^{-1}, which exhibited multi-vibrational bands. The spectrum showed bands at 493, 764, 1038, 1105, 1251,

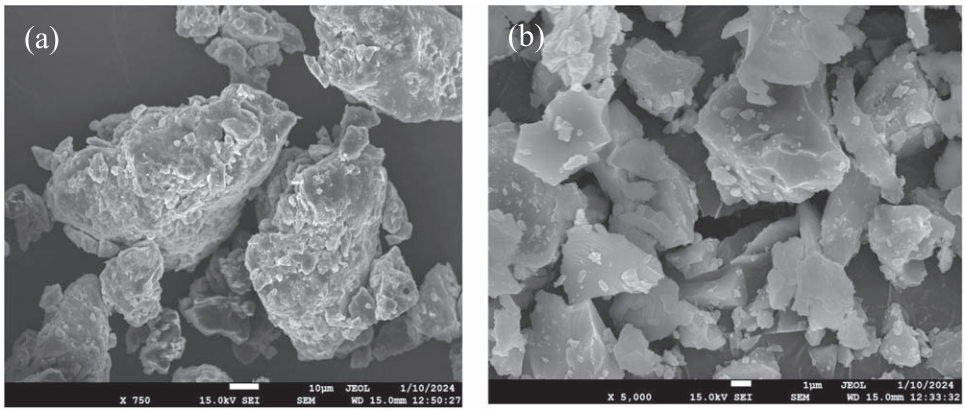


Figure 2. SEM image recorded For (a) an un-doped (b) 4.0 mol% of Dy^{3+} ion doped $KCa(PO_3)_3$.

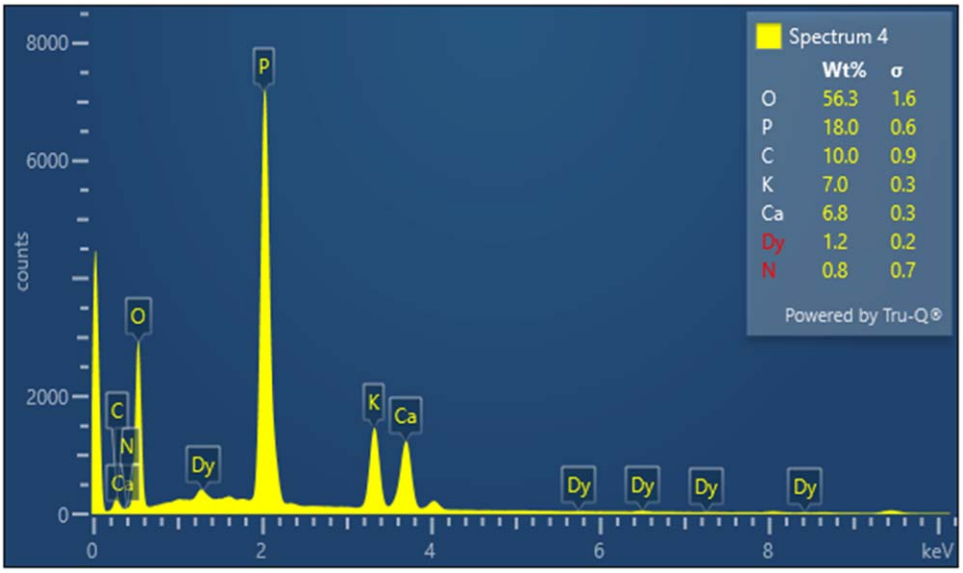


Figure 3. EDX image ofof $KCa(PO_3)_3$: xDy^{3+} (x = 4 mol%) phosphor.

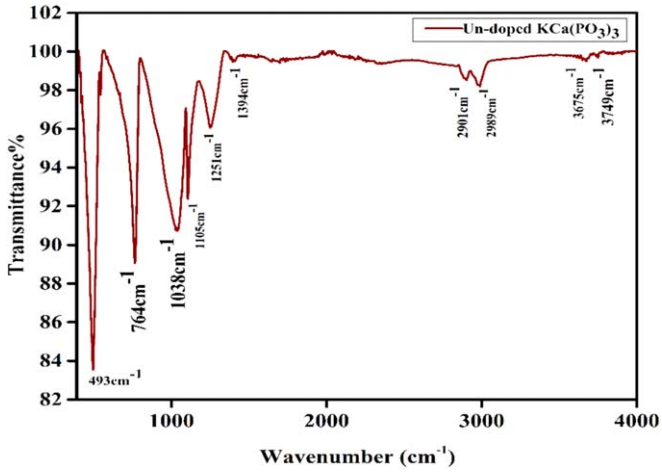


Figure 4. FT-IR spectrum of un-doped $KCa(PO_3)_3$ phosphor.

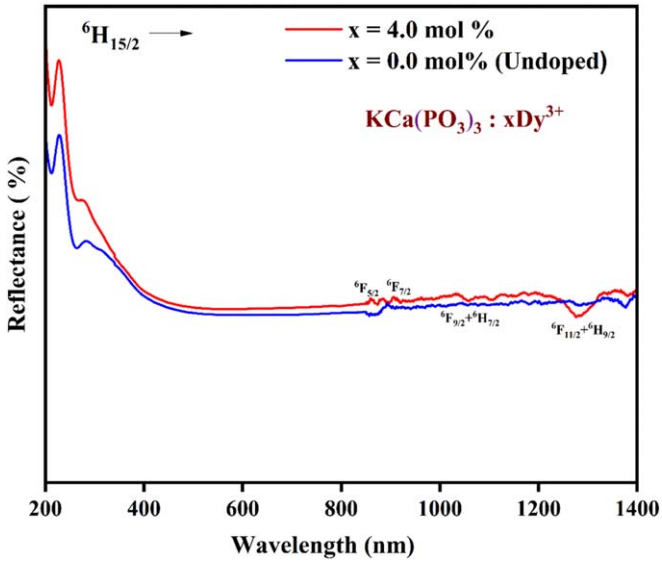


Figure 5. Diffuse reflectance spectra of undoped and $KCa(PO_3)_3$: xDy^{3+} (x = 4 mol%) phosphor.

1394, 2901, 2989, 3675, and 3749 cm^{-1}. The stretching vibrations of the PO_2 were recorded in the 1000–1200 cm^{-1} and 1200–1400 cm^{-1} ranges, for the symmetric and asymmetric stretching, respectively, while the bending vibrations were detected in the range of 400–550 cm^{-1}. The vibrations of the P-O-P oxygen bridges were detected in the region of asymmetric mode (1000–1050 cm^{-1}) and symmetric mode (600–800 cm^{-1}) [21].

The measured FT-IR spectra matched well with the theoretical considerations presented earlier. The observed bending vibrational wavenumbers correspond to specific normal modes, with the asymmetric stretching (O-P-O) showing a strong vibration at 1251 cm^{-1} and the symmetric stretching (O-P-O) exhibiting a very strong vibration at 1105 cm^{-1}. Similarly, the asymmetric stretching (P-O-P) showed a very strong vibration at 1038 cm^{-1}, and the symmetric stretching (P-O-P) had a strong vibration at 764 cm^{-1}. The antisymmetric stretching of bonded water molecules and the presence of hydroxyl radicals were indicated by the peaks observed at 2901 cm^{-1} and 2989 cm^{-1}, respectively. The band observed at 3675 cm^{-1} was responsible for the stretching peak of the hydroxyl group [23]. The multi-component patterns of these spectra showed the same number of bands as those reported in previous works for ABP_3O_9-type compounds.

3.4. Diffuse reflectance spectra (DRS) analysis

Phosphors utilized in w-LEDs have an optical band gap (E_g), a crucial parameter. The DRS of undoped and KCP:xDy^{3+}, x = 4 mol% is shown in figure 5 between 200 and 1400 nm range [24]. The DRS spectral data were converted into the Kubelka–Munk function, and the optical band gap was then calculated using the following equation [25].

$$F(R) = \frac{1 - R^2}{2R} = \frac{\alpha}{S} \tag{1}$$

Here, R stands for sample reflectance, S for scattering coefficient, and α for absorption coefficient. The equation shows the following relationship between energy bandgap (E_g) and absorption coefficient (α).

$$\alpha h\nu = C(h\nu - E_g)^n \tag{2}$$

Here, n can have values of 2 or 1/2, corresponding to indirect and direct transitions, respectively, depending on the kind of transition. C is an energy-independent constant (in the current case, = 1). The unit of photon energy is hν. Since α is exactly proportional to F(R), the above equation can be expressed in this way:

$$F(R)h\nu = C(h\nu - E_g)^n \tag{3}$$

To determine the band gap for the KCP:Dy^{3+} phosphor, a tauc plot was created b/w hν and $[F(R)h\nu]^{1/2}$ [24]. The optical band gap (E_g) can be derived by extrapolating the linear portion of the plot. The plot is shown in figure 6 and the calculated E_g values for undoped and optimised KCP phosphors are 4.60 and 4.49 eV respectively.

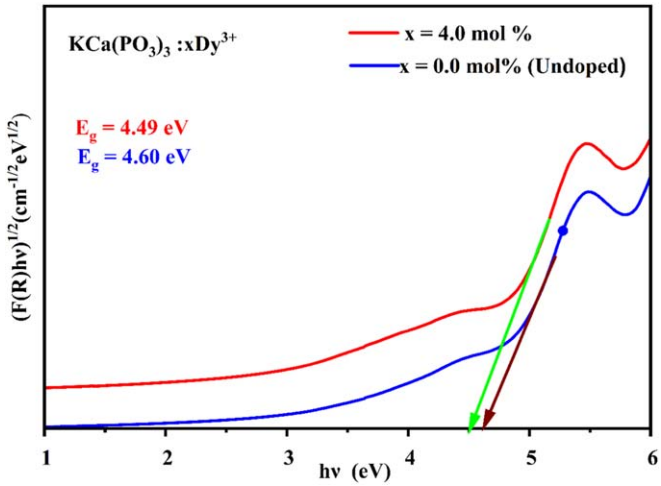


Figure 6. Tauc plot for direct optical band gap of undoped and $KCa(PO_3)_3$: xDy^{3+} (x = 4 mol%) phosphor.

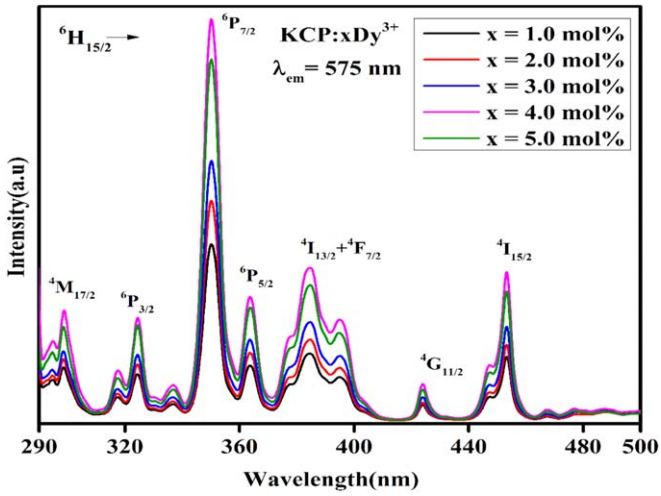


Figure 7. Excitation spectra of Dy^{3+} ions doped $KCa(PO_3)_3$ phosphors under 575 nm emission wavelength.

3.5. Photoluminescence spectral analysis (PL) of $KCa(PO_3)_3$

3.5.1. Excitation spectra

The excitation spectra of KCP: xDy^{3+} (x = 1, 2, 3, 4, 5 mol%) phosphors were shown in figure 7. Several sharp excitation peaks have been detected within a specific range of wavelengths (290–500 nm). There are 5 primary peaks of the excitation spectra that are caused by the f-f electronic transitions of Dy^{3+} ions, with the maximum excitation wavelength situated at 350 nm. The electronic transitions responsible for these peaks are $^6H_{15/2}$ to $^4M_{17/2}$ [298 nm], ($^6P_{3/2}$) [324 nm], $^6P_{7/2}$ [350 nm], $^6P_{5/2}$[364 nm], ($^4I_{13/2}$+$^4F_{7/2}$) [386 nm], and $^4I_{15/2}$[453 nm] [26]. The KCP: Dy^{3+} phosphor is suitable for excitation by near-UV LEDs, as indicated by the main excitation bands at 350–400 nm [27].

3.5.2. Emission spectra

The emission spectra of KCP: xDy^{3+} (x = 1, 2, 3, 4, 5 mol%) shown in figure 8 were examined under 350 nm excitation. We found that the emission shape and position didn't change significantly with the increase in the concentration of Dy^{3+} doping, except for the emission intensity. The spectra shows two strong characteristic peaks; at 482 nm corresponding to blue emission and 575 nm corresponding to yellow emission, which were attributed to the transition from $^4F_{9/2}$ to $^6H_{15/2}$ and $^6H_{13/2}$, respectively and weak peaks observed at 665 nm (red emission) ascribed the transition $^4F_{9/2}$ to $^6H_{11/2}$. Our observations suggest that the optical characteristics of KCP: Dy^{3+} are directly impacted by the configuration of the host material in which they are embedded. When the Dy^{3+} ion is positioned at a site with high symmetry, i.e. a site that has an inversion center, the yellow emission

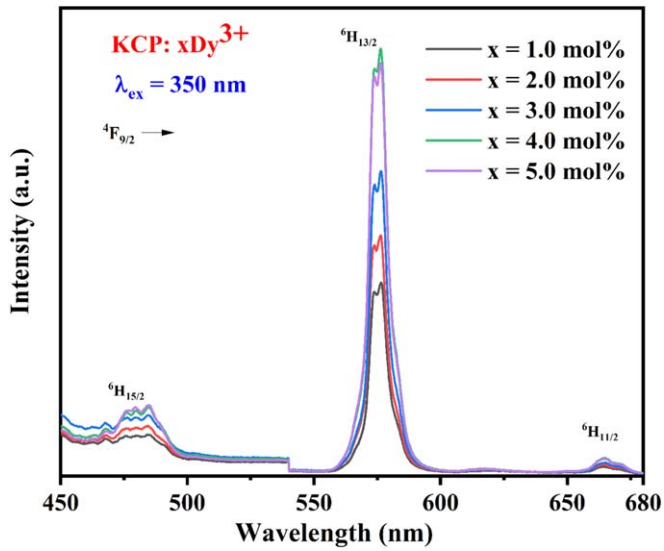


Figure 8. Emission spectra of Dy^{3+} ions doped $KCa(PO_3)_3$ phosphors under 350 nm excitation wavelength.

is the predominant outcome. However, when the Dy^{3+} ion is positioned at a position without a center of symmetry, which is typically a site exhibiting low symmetry, its emission intensity is observed to be considerably amplified [28]. The emission of yellow light with a wavelength of 575 nm ($^4F_{9/2}$-$^6H_{13/2}$) is a result of a forced dipole transition (electric transition). This type of transition happens only at low symmetries without an inversion center. The intensity of this emission is highly dependent on the crystal-field environment. We found that its intensity increases immensely as a function of Dy^{3+} concentration. The blue band's intensity corresponding to the 482 nm ($^4F_{9/2}\rightarrow{}^6H_{15/2}$) dipole transition, (which refers to the change in the magnetic moment during a transition between energy levels), is known to exhibit a relatively low sensitivity to the effects of crystal fields. The analysis of emission spectra suggests that the concentration of Dy^{3+} has a significant impact on the intensity of the blue transition $^4F_{9/2}\rightarrow{}^6H_{15/2}$. Specifically, an increase in Dy^{3+} concentration up to 4% results in increased emission intensity. But beyond this concentration, there is a discernible drop in emission intensity. The strength of the $^4I_{13/2}\rightarrow{}^6H_{15/2}$, $^4I_{15/2}\rightarrow{}^6H_{15/2}$, and $^4F_{9/2}\rightarrow{}^6H_{11/2}$ transitions, however, rapidly dropped as the concentration of Dy^{3+} increased [29–31]. The observed evolution is a result of the concentration quenching phenomenon and the impact of population on the $^4I_{15/2}$ and $^4I_{13/2}$ energy levels. The amplified resonant energy transfer from the excited energy state to the low energy or ground state ($^4F_{9/2}\rightarrow{}^6H_{15/2}$) of the nearby Dy^{3+} ion, as well as the cross-relaxation phenomenon between the excited Dy^{3+} ion (donor) and the ground Dy^{3+} ion (acceptor), are responsible for the increase in quenching intensities with increasing Dy^{3+} ion concentration.

At 4 mol% concentration of Dy^{3+}, concentration quenching occurred. However, when the concentration of these ions becomes too high, the possibility of energy transfer between ions rises. As a result, more excitation energy is transferred via non-radiative processes, resulting in decreased luminescence efficacy. To examine the energy transmission mechanism between Dy^{3+} ions, we computed the critical distance. The Blasse theory equation may be used to determine the critical radius distance R_c [23, 32].

The Blasse equation is:

$$R_c \approx 2\left[\frac{3V}{4\pi X_c N}\right]^{\frac{1}{3}} \tag{4}$$

Where N is the number of cations per unit cell (N = 2), X_c is the threshold concentration of Dy^{3+} ($X_C = 0.04$), and V is the volume per unit cell ($V = 416.33^{*}10^4$ Å).

The critical distance (R_c) between activator ions in $KCa(PO_3)_3$:xDy^{3+} was 21.50 Å, indicating that non-radiative energy transfer predominantly occurs through exchange contacts or multilayer interactions. For energy transfer to occur, the activator and sensitizer need to be positioned close to each other within the lattice. Additionally, there should be significant overlap in their wave functions, which allows for electron exchange to take place. The critical distance for this exchange is less than 5 Å, but the calculated R_c values were much higher, indicating that energy transfer between Dy^{3+}-Dy^{3+} ions in $KCa(PO_3)_3$:xDy^{3+} happens through multilevel interactions. According to the Dexter theoretical model:

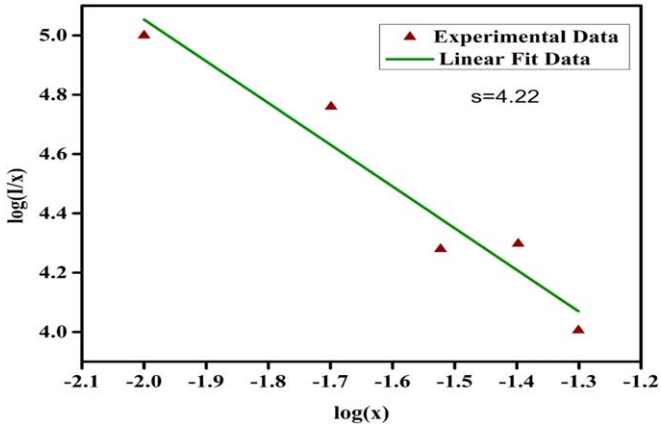


Figure 9. Plot of log (I/x) with log(x).

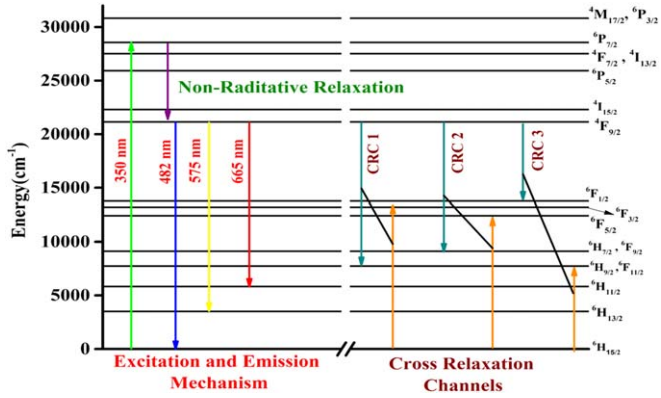


Figure 10. Partial energy level diagram and possible cross-relaxation channels of Dy^{3+} ion doped $KCa(PO_3)_3$ phosphor.

$$\log\left(\frac{I}{x}\right) = A - \frac{s\log(x)}{3} \tag{5}$$

Where A signifies a constant, d represents the mixture's dimensions (= 3) and I is the emission intensity [33, 34]. We use the value of s to identify the type of multilevel interactions. The types of interactions, including closest neighbor ion, dipole–dipole, dipole-quadrupole, and quadrupole-quadrupole interactions, are quantified by their respective multipole moments. Specifically, these interactions are represented by the values of 3, 6, 8, and 10 for the corresponding multipole moments respectively. To study the values of s, we select samples with x = 0.01, 0.02, 0.03, 0.04, and 0.05. Subsequently, we determine the emission intensity and generate a linear graph of KCP:xDy^{3+} that is shown in figure 9 between log(I/x) versus log(x). The (s/3) value is determined by analyzing the gradient of the linear plot b/w log(I/x) and log(x). Upon analyzing the data, we found that the slope is −1.40712. This value helped us determine that the value of s is 4.2214, which is less than 6. This suggests that energy transfer due to nearest neighbouring ions is the primary mechanism causing concentration quenching in the sample [35].

Asymmetric molecules exhibit inherent dipoles because of the unequal distribution of electrons around the positively charged nucleus. The dipole–dipole interaction's dimension is determined by the two dipoles' center distance and relative orientation. The interaction is computed using the potential energy of each dipole. Concentration quenching results from a substantial energy exchange between activators that happens when the concentration of activators reaches a certain level [36]. The schematic energy level diagram shown in figure 10 illustrates the possible electronic transitions of the Dy^{3+} in the $KCa(PO_3)_3$ polyphosphate phosphor [37]. The higher energy level ($^6P_{7/2}$) of the Dy^{3+} ions becomes populated by Dy^{3+} excitation (350 nm). Due to inadequate energy difference between the states to produce a radiative transition, the ions undergo non-radiative transitions (phonon relaxation) from the excited level to the ground state ($^4F_{9/2}$). Ultimately descends to the various ground-level states ($^6H_{15/2}$, $^6H_{13/2}$, and $^6H_{11/2}$) to release blue, yellow, and red photons, which are associated with the Dy^{3+} ion transitions at $^4F_{9/2} \rightarrow {}^6H_{15/2}$, $^4F_{9/2} \rightarrow {}^6H_{13/2}$, and $^4F_{9/2} \rightarrow {}^6H_{11/2}$ [18].

This energy exchange reduces the amount of luminous energy levels and lessens the intensity of the fluorescence. The three cross-relaxation processes of Dy^{3+} are as follows [36]:

$$\text{CRC 1:} \quad ({}^4F_{9/2} + {}^6H_{15/2} \rightarrow ({}^6H_{9/2}, {}^6F_{11/2}) + {}^6F_{3/2})$$

An initial Dy^{3+} ion descends from level $^4F_{9/2}$ to the intermediate level ($^6H_{9/2}$, $^6F_{11/2}$), while another Dy^{3+} ion is excited from the ground state level $^6H_{15/2}$ to level $^6F_{3/2}$.

$$\text{CRC 2:} \quad ({}^4F_{9/2} + {}^6H_{15/2} \rightarrow ({}^6H_{7/2}, {}^6F_{9/2}) + {}^6F_{5/2})$$

An initial Dy^{3+} ion descends from level $^4F_{9/2}$ to the intermediate level $^6H_{7/2}$+$^6F_{9/2}$, while another Dy^{3+} ion is excitedfrom the ground state level $^6H_{15/2}$ to level $^6F_{5/2}$.

$$\text{CRC 3:} \quad ({}^4F_{9/2} + {}^6H_{15/2} \rightarrow {}^6F_{1/2} + ({}^6H_{9/2}, {}^6F_{11/2}))$$

An initial Dy^{3+} ion descends from level $^4F_{9/2}$ to the intermediate level $^6F_{1/2}$, while another Dy^{3+} ion is excited from the ground state level $^6H_{15/2}$ to level $^6H_{9/2}$+$^6F_{11/2}$.

3.6. Y/B ratio

The luminescence intensity and the site symmetry of RE ions are generally inversely related. Solid-state lighting appliances prioritize the Dy^{3+} ion because it emits the main colors that are necessary for the creation of white light [38]. In general, white light can be produced by precisely balancing blue and yellow light. As seen in figure 8, the yellow band's intensity (at 575 nm) is comparatively higher than the blue band's intensity (at 350 nm). The production of white light is significantly influenced by the intensity ratios of the blue and yellow bands, which are the ratios of their combined intensities [39]. The concentration of dopant (in this case, Dy^{3+}) ions, the host phosphor matrix, and the excitation source all have an impact on the yellow-to-blue (Y/B) intensity ratio. The Y/B intensity ratio is significantly impacted by the inversion symmetry and local coordination surrounding nearby Dy^{3+} ions. Moreover, it is found that the local coordinated regions around Dy^{3+} ions are the least symmetric with respect to neighboring atoms in the absence of inversion symmetry. Due to the increased covalency, the emission spectrum exhibits an intense yellow band as opposed to a blue one. Y/B intensity ratios are comparatively more than 1, indicating that light is being generated in the white area. The higher Y/B ratios support the high covalent nature of Dy^{3+} ions. Consequently, the KCP:Dy^{3+} 4.0 mol in the as-prepared KCP phosphors shows the highest Y/B value and, consequently, the maximum covalency.

3.7. CIE Co-ordinate

To understand the colorimetric performance of $KCa(PO_3)_3$: xDy^{3+} ions (x = 1.0, 2.0, 3.0, 4.0, 5.0 mol %) phosphor, the 'Commission Internationale de l'Eclairage' (CIE) chromaticity coordinates at $\lambda_{ex} = 350$ nm were acquired from the emission spectrum data. The figure 11 displays the calculated CIE coordinates for every as-synthesized $KCa(PO_3)_3$ phosphor at $\lambda_{ex} = 350$ nm and $\lambda_{em} = 575$ nm. The coordinates of the as-synthesized Dy^{3+} doped $KCa(PO_3)_3$ phosphors, as indicated in the CIE diagram lie near the central white region. This indicates that $KCa(PO_3)_3$ phosphors doped with Dy^{3+} ions are suitable for solid-state lighting and plasma display panels [40]. Another important component in colorimetric computations used to look into color saturation is color purity (CP) [41]. The estimation of CP can be done using the formula below.

$$Color\ \ purity = \frac{\sqrt{(x - x_{ee})^2 + (y - y_{ee})^2}}{(x_d - x_{ee})^2 + (y_d - y_{ee})^2} \tag{6}$$

where (x, y) represents the color coordinates of the $KCa(PO_3)_3$ phosphor, (x_{ee}, y_{ee}) stands for equal energy point, and (x_d, y_d) indicates points of dominant wavelength. The Dy^{3+} (x = 4.0 mol%) ions doped $KCa(PO_3)_3$ phosphor is found to have 16% color purity under 350 nm excitation. When triggered by blue or n-UV light, Dy^{3+} (x = 4.0 mol%) ions doped $KCa(PO_3)_3$ phosphor can be used as a white emitting component in w-LEDs.

When determining whether a light source is warm or cold, one of the most crucial parameters is the correlated color temperature (CCT) [42]. CCT is determined using the equation below.

$$CCT = -449n^3 + 525n^2 + 6823.3n + 5520.32 \tag{7}$$

The color coordinates are n = (x-0.332)/(0.1858-y). The prepared phosphor is a cool emission source, as indicated by the assessed CCT values as mentioned in table 1. The calculated CCT values of the KCP:xDy^{3+} phosphors are significantly more than 4000 K, indicating that cool white light can be created from these phosphors.

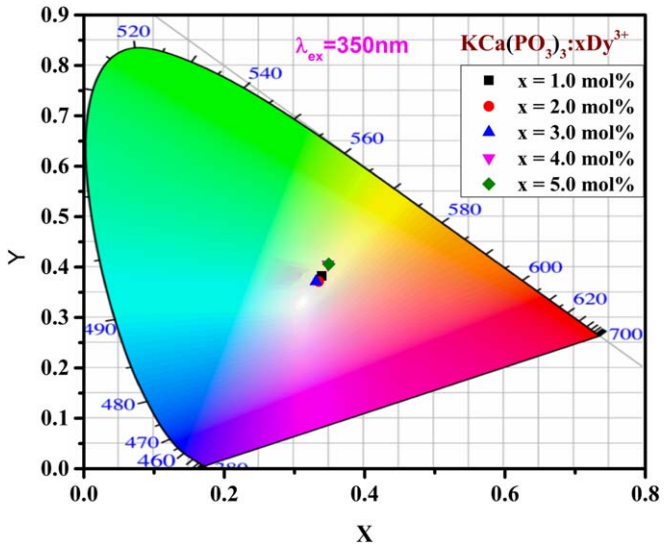


Figure 11. CIE chromaticity coordinates of Dy^{3+} ions in $KCa(PO_3)_3$ phosphors.

Table 1. CIE coordinates, CCT values of KCP: xDy^{3+} (x = 1.0, 2.0, 3.0, 4.0, 5.0 mol%) under 350 nm excitation wavelength.

Sample ID KCP: xDy^{3+}	CIE coordinates (x, y)	CCT values (K)
x = 1.0 mol%	(0.3389, 0.3823)	5287.127
x = 2.0 mol%	(0.3346, 0.3712)	5428.813
x = 3.0 mol%	(0.3307, 0.3715)	5570.698
x = 4.0 mol%	(0.3678, 0.4051)	4497.287
x = 5.0 mol%	(0.368, 0.4058)	5056.40

3.8. Photoluminescence (PL) decay

The PL decay patterns under room temperature $\lambda_{ex} = 350$ nm and $\lambda_{em} = 575$ nm for $KCa(PO_3)_3$: x Dy^{3+} (x = 1.0, 2.0, 3.0, 4.0, and 5.0 mol%) ions phosphor are represented in figure 12. The decay curves were fitted with various exponential equations; however, for all concentrations of Dy^{3+} ions, the bi-exponential equation showed the best fit. This suggests that dopant ions are interacting significantly, leading to energy transfer from excited Dy^{3+} ions (donor) to unexcited Dy^{3+} ions (acceptor) and that the RE ions in the phosphor lattice are probably going to acquire two different crystallographic sites [43, 44].

The following equation can be used to represent the luminous intensity:

$$I = I_0 + A_1 \exp\left(\frac{-t}{\tau_1}\right) + A_2 \exp\left(\frac{-t}{\tau_2}\right) \tag{8}$$

Where I_o and I represent, respectively, the luminescence intensities at time t = 0 and some other point in time. A_1 and A_2 are the fit constants for bi-exponential fitting, and τ_1 and τ_2 are the components for rapid and slow decay, respectively [45]. When modeling a bi-exponential curve, the average lifetime values are provided by:

$$\tau_{exp} = \frac{A_1\tau_1^2 + A_2\tau_2^2}{A_1\tau_1 + A_2\tau_2} \tag{9}$$

The as-synthesized $KCa(PO_3)_3$: xDy^{3+} (x = 1.0, 2.0, 4.0, and 5.0 mol%) phosphor have τ_{exp} values of 377.74, 327.04, 314.21, 308.84, and 293.03 μs, respectively. It was observed that when the quantity of Dy^{3+} ions in the phosphors was increased, the values of τ_{exp} decreased.

The decrement in a lifetime values with the increase inthe concentration of Dy^{3+} ions was further examined by Auzel's model utilizing equation given below [46]:

$$\tau = \frac{\tau_0}{1 + \frac{c}{c_0}e^{-N/3}} \tag{10}$$

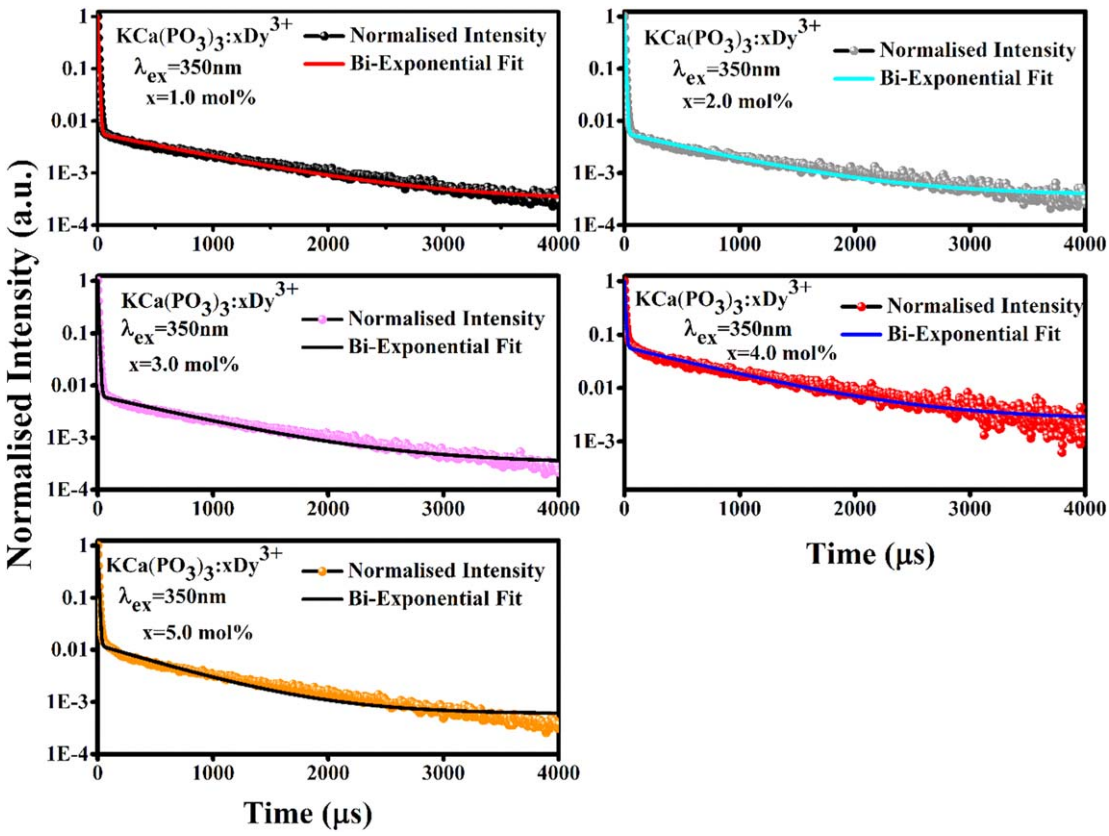


Figure 12. PL decay profiles of Dy^{3+} ions doped $KCa(PO3)_3$ phosphors at $\lambda_{ex} = 350$ nm wavelength.

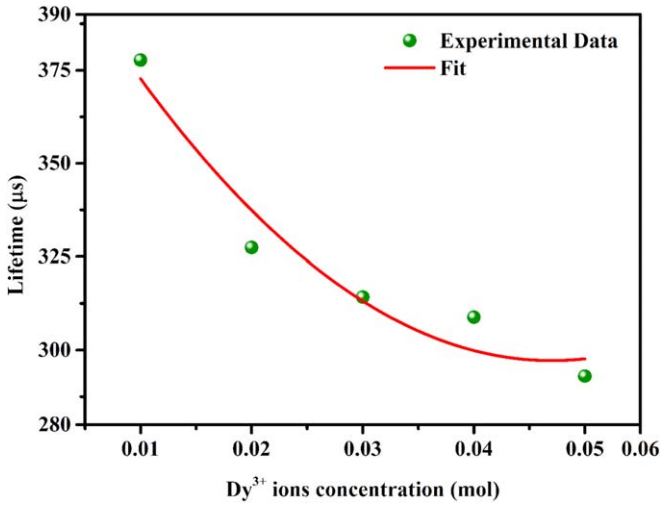


Figure 13. Dependence of photoluminescence decay lifetime of the $^4F_{9/2}$ state on the doping concentration of Dy^{3+} based on Auzel's model.

where τ represents the calculated lifetime, τ_0 represents the intrinsic radiative lifetime, c stands for concentration of Dy^{3+} ions, c_0 represents a constant and N denotes the number of phonons generated during the relaxation. Figure 13 demonstrates that the Dy^{3+} decay lifetime in KCP phosphor fits well with Auzel's model and the value of τ_0 from the fitting is 419.07 μs.

The quantum efficiency (η) is computed using intrinsic lifetime value as follows [47, 48].

$$\eta = \frac{\tau}{\tau_0} = \frac{A_R}{A_R + A_{nR}} \tag{11}$$

Table 2. Photoluminescence decay lifetime, non-radiative rate and quantum efficiency of the KCP : xDy^{3+} (x = 1.0, 2.0, 4.0, and 5.0 mol%) phosphors.

Sample ID KCP: xDy^{3+}	Decay Lifetime (μs)	Non radiative rate (s^{-1})	Quantum Efficiency (%)
x = 1.0 mol%	377.74	261.0874	90.13769
x = 2.0 mol%	327.04	671.4629	78.04026
x = 3.0 mol%	314.21	796.3153	74.97871
x = 4.0 mol%	308.84	851.6643	73.69702
x = 5.0 mol%	293.03	1026.43	69.92292

$$\frac{1}{\tau} = \frac{1}{\tau_0} + A_{nR} \tag{12}$$

here the rate of relaxation by non-radiative and radiative mechanisms are denoted by A_{nR} and A_R, respectively. Table 2 shows quantum efficiency values (%) along with decay lifetime and non-radiative rates (s^{-1}) for $^4F_{9/2}$ state.

4. Conclusions

In our work, Dy^{3+} doped $KCa(PO_3)_3$ phosphors were successfully synthesized via conventional solid-state reaction technique and were investigated their structural, morphological, and PL characteristics. First, the phase purity of the produced samples was confirmed by analyzing the XRD patterns. $KCa(PO_3)_3$: Dy^{3+} formed a hexagonal structure when it crystallized with a P-6c2(188) space group. Using SEM, morphological analysis was performed, and particle sizes were determined to be in the micrometre (3–5 μm) range. Multiple molecular bonds that are present in the structure are shown by the FT-IR spectra. The band gap values, 4.60 & 4.49 eV for undoped and 4.0 mol% of Dy^{3+}, were determined by linear fitting of the tauc plot of Dy^{3+} in the $KCa(PO_3)_3$ host. The excitation spectra show that the previously stated phosphors can be efficiently excited by n-UV light. A number of experiments were used to characterize the samples' luminescence properties. The PL emission spectra findings demonstrated that the samples generated white light with two prominent peaks at 482 nm (blue band) and 575 nm (yellow band) upon excitation by 350 nm UV radiation. The electronic transitions of dysprosium ions at $^4F_{9/2} \rightarrow {}^6H_{15/2}$ and $^4F_{9/2} \rightarrow {}^6H_{13/2}$ were matched by these peaks, respectively. Concentration quenching was observed for Dy^{3+} doping at a concentration of 4.0% mole. The main cause of this quenching was found to be dipole–dipole energy transfer. The computed CIE chromaticity coordinates for $KCa(PO_3)_3$ for 4.0 mol% of the Dy^{3+} phosphor lie in the deep white area. The PL decay profile displays bi-exponential behavior with decay times being in microseconds (μs) range. As the concentration of Dy^{3+} ions increases, the computed τ_{exp} values decrease because of the transfer of energy between Dy^{3+} - Dy^{3+} ions. The results of the aforementioned studies on the as prepared KCP:Dy^{3+} phosphors highlight that these materials can be used for white light generating LEDs and other photonic devices applications.

Acknowledgments

One of the authors, Ms. Sheetal Kumari is very much thankful to CSIR-UGC for the award of a Junior Research Fellowship (JRF) (UGC Ref. 1431/CSIR-UGC NET JUNE 2019). We are really grateful to Dr. Manju Srivastava and Dr. Shyama Prasad for the SEM facility provided by the Department of Chemistry at Dayalbagh Educational Institute (DEI), Deemed University, Agra.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Author contributions

Deepti: Experimentation, data analysis, manuscript writing.
Sandip Maurya: Experimentation, data analysis, manuscript writing.
Sheetal Kumari: Work plan, supervision, review, data analysis.

Pooja Rohila: Work plan, supervision, review, data analysis.
Aman Prasad: Review, data analysis.
A S Rao: Workplan, supervision, review

Data availability

The data that support the findings of this study are available on request from the corresponding author Prof. A.S. Rao. The data are not publicly available due to restrictions, e.g. their containing information that could compromise the privacy of research participants.

Declaration

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ORCID iDs

A S Rao https://orcid.org/0000-0001-8199-6505

References

[1] Yim D K, Song H J, Roh H S, Kim S J, Han B S, Jin Y H, Hong H S, Kim D W and Hong K S 2013 Luminescent properties of $RbSrPO_4:Eu^{2+}$ phosphors for near-UV-based white-light-emitting diodes *Eur. J. Inorg. Chem.* **26** 4662–6
[2] Balakrishna A and Ntwaeaborwa O M 2017 Study of luminescent behavior and crystal defects of different $MNa[PO_4]$-Dy^{3+} phosphors (M = Mg, Ca, Sr and Ba) *Sensors Actuators, B Chem.* **242** 305–17
[3] Narendran N, Bullough J D, Maliyagoda N and Bierman A 2001 What is useful life for white light leds? *J. Illum. Eng. Soc.* **30** 57–67
[4] Ohno Y 2005 Spectral design considerations for white LED color rendering *Opt. Eng.* **44** 111302
[5] Munirathnam K, Nagajyothi P C, Prakashbabu D, Deva Prasad Raju B and Shim J 2019 x-ray photoelectron spectroscopy and optical analysis of pure white light emitting Dy^{3+} and Mn^{2+} codoped $Na_3Y(PO_4)_2$ phosphors for solid-state lighting *Ceram. Int.* **45** 686–94
[6] Muthu S, Schuurmans F J and Pashley M D 2002 Red, green, and blue LED based white light generation: issues and control *Conf. Rec. - IAS Annu. Meet. (IEEE Ind. Appl. Soc.* **1** 327–33
[7] Bando K, Sakano K, Noguchi Y and Shimizu Y 1998 Development of High-bright and Pure-white LED Lamps *J. Light Vis. Environ.* **22** 2–5
[8] Chemingui S, Ferhi M, Horchani-Naifer K and Férid M 2015 Synthesis and luminescence characteristics of Dy^{3+} doped $KLa(PO_3)_4$ *J. Lumin.* **166** 82–7
[9] Krishna Bharat L, Jeon Y I and Yu J S 2014 Synthesis and luminescent properties of trivalent rare-earth (Eu^{3+}, Tb^{3+}) ions doped nanocrystalline $AgLa(PO_3)_4$ polyphosphates *J. Alloys Compd.* **614** 443–7
[10] Kim E D, Do Y R, Oh I H and Huh Y D 2021 Preparation and photoluminescence properties of transparent suspensions of $Ca(Y_{1-x}Eu_x)_2(MoO_4)_4$ nanophosphors *Opt. Mater. (Amst).* **119** 111394
[11] Ngee H L, Hatsumori T, Uematsu K, Ishigaki T, Toda K and Sato M 2009 Synthesis of phosphate phosphor for a white LED *Phys. Procedia* **2** 171–83
[12] Pawade V B, Zanwar A, Birmod R P, Dhoble S J and Koao L F 2017 Optical and bandgap study of rare earth doped phosphate phosphor *J. Mater. Sci., Mater. Electron.* **28** 16306–13
[13] Guo Z, Zhu Z, Zhang X and Zhou L 2018 Photoluminescent properties and Judd-Ofelt analysis of Eu^{3+}-doped $NaMg_{(PO3)_3}$ red phosphor *J. Lumin.* **202** 484–8
[14] Chand S, Mehra R and Chopra V 2022 Recent advancements in calcium based phosphate materials for luminescence applications *J. Lumin.* **252** 119383
[15] Nandanwar C M, Kokode N S, Yerpude A N and Dhoble S J 2023 Luminescence properties of $LaPO_4$:RE (RE = Dy^{3+}, Eu^{3+}, Sm^{3+}) orthophosphate phosphor for n-UV solid-state lighting prepared by wet chemical synthesis *J. Mater. Sci., Mater. Electron.* **34** 1–8
[16] Bedyal A K, Kumar V and Swart H C 2020 Influence of an adjoining cation on the luminescence performance of the Dy^{3+} doped $A_3Gd(PO_4)_2$; (A = Na, K) phosphors *J. Alloys Compd.* **845** 156352
[17] Ferhi M, Toumi S, Horchani-Naifer K and Ferid M 2017 Single phase $GdPO_4:Dy^{3+}$ microspheres blue, yellow and white light emitting phosphor *J. Alloys Compd.* **714** 144–53
[18] Bedyal A K, Kunti A K, Kumar V and Swart H C 2019 Effects of cationic substitution on the luminescence behavior of Dy^{3+} doped orthophosphate phosphor *J. Alloys Compd.* **806** 1127–37
[19] Zhang L and Han G C 2013 Novel Dy^{3+}-doped $Gd(PO_3)_3$ white-light phosphors under VUV excitation for Hg-free lamps application *Chinese Phys. B* **22** 027803
[20] Wang S, Xu C and Qiao X 2020 Enhanced photoluminescence of red-emitting $KCa(PO_3)_3$: Eu^{3+} phosphors by charge compensation *Opt. Mater. (Amst).* **107** 110102
[21] Watras A, Matraszek A, Godlewska P, Szczygieł I, Wojtkiewicz J, Brzostowski B, Banach G, Hanuza J and Dereń P J 2014 The role of the Ca vacancy in the determination of the europium position in the energy gap, its valence state and spectroscopic properties in $KCa(PO_3)_3$ *Phys. Chem. Chem. Phys.* **16** 5581–8
[22] Krishnapriya T, Jose A, Anna Jose T, Joseph C, Unnikrishnan N V and Biju P R 2021 Luminescent kinetics of Dy^{3+} doped $CaZn_2(PO_4)_2$ phosphors for white light emitting applications *Adv. Powder Technol.* **32** 1023–32
[23] Sahu M K and Mula J 2019 White light emitting thermally stable bismuth phosphate phosphor $Ca_3Bi(PO_4)_3:Dy^{3+}$ for solid-state lighting applications *J. Am. Ceram. Soc.* **102** 6087–99

[24] Kumari S, Rao A S and Sinha R K 2023 Structural and photoluminescence properties of Sm^{3+} ions doped strontium yttrium tungstate phosphors for reddish-orange photonic device applications *Mater. Res. Bull.* **167** 112419
[25] Mahajan R, Kumar S, Prakash R, Kumar V, Choudhary R J and Phase D M 2019 x-ray photoemission and spectral investigations of Dy^{3+} activated magnesium pyrophosphate phosphors *J. Alloys Compd.* **777** 562–71
[26] Mahajan R, Prakash R, Kumar S, Kumar V, Choudhary R J and Phase D M 2021 Surface and luminescent properties of $Mg_3(PO_4)_2$:Dy^{3+} phosphors *Optik (Stuttg).* **225** 165717
[27] Vidyadharan V, Sreeja E, Jose S K, Joseph C, Unnikrishnan N V and Biju P R 2016 Spectroscopic and photoluminescence characterization of Dy^{3+} in $Sr_{0.5}Ca_{0.5}TiO_3$ phosphor *Luminescence* **31** 202–9
[28] Shinde K N, Dhoble S J and Kumar A 2011 Synthesis of novel Dy^{3+} activated phosphate phosphors for NUV excited LED *J. Lumin.* **131** 931–7
[29] Shivaramu N J, Nagabhushana K R, Lakshminarasappa B N and Singh F 2016 Ion beam induced luminescence studies of sol gel derived Y_2O_3:Dy^{3+} nanophosphors *J. Lumin.* **169** 627–34
[30] Poma P Y, Santos W Q, Sales T O, Gouveia-Neto A S and Jacinto C 2017 Tunable light emission mediated by energy transfer in Tm^{3+}/Dy^{3+} co-doped LaF_3 nanocrystals under UV excitation *J. Lumin.* **188** 18–23
[31] Wang Y, Liu X, Jing L and Niu P 2016 Tunable white light and energy transfer of Dy^{3+} and Eu^{3+} doped $Y_2Mo_4O_{15}$ phosphors *Ceram. Int.* **42** 13004–13010
[32] Fawad U, Kim H J, Khan S, Khan M and Ali L 2016 Photoluminescent properties of white-light-emitting $Li_6Y(BO_3)_3$:Dy^{3+} phosphor *Solid State Sci.* **62** 1–5
[33] Zhao D, Zhang S R, Fan Y P, Liu B Z and Zhang R J 2020 Thermally stable phosphor $KBa_2(PO_3)_5$:$Eu^{2+}$with broad-band cyan emission caused by multisite occupancy of Eu^{2+} *Inorg. Chem.* **59** 8789–99
[34] Wang J, Wang J and Duan P 2013 Luminescent properties of Dy^{3+} doped $Sr_3Y(PO_4)_3$ for white LEDs *Mater. Lett.* **107** 96–8
[35] Faoro R, Moglia F, Tonelli M, Magnani N and Cavalli E 2009 Energy levels and emission parameters of the Dy^{3+} ion doped into the YPO_4 host lattice *J. Phys. Condens. Matter* **21** 275501
[36] Tang W, Guo Q, Su K, Liu H, Zhang Y, Mei L and Liao L 2022 Structure and photoluminescence properties of Dy^{3+} doped phosphor with whitlockite structure *Materials (Basel).* **15** 4–13
[37] Carnall W T, Fields P R and Rajnak K 1968 Electronic energy levels of the trivalent lanthanide aquo ions. IV. Eu^{8+} *J. Chem. Phys.* **49** 4424–42
[38] Huang W, Liu Q, Li X, Zhu Q, Wang L and Zhang Q 2019 Dy^{3+}-doped $BaLaMgSbO_6$ double perovskite highly efficient white phosphor *Ceram. Int.* **45** 15624–15628
[39] Vijayakumar R, Venkataiah G and Marimuthu K 2015 Structural and luminescence studies on Dy^{3+} doped boro-phosphate glasses for white LED's and laser applications *J. Alloys Compd.* **652** 234–43
[40] Verma T and Agrawal S 2018 Optical studies and estimation of kinetic parameters for dysprosium activated yttrium oxide phosphors *Optik (Stuttg).* **160** 361–70
[41] Su Q, Liang H, Li C, He H, Lu Y, Li J and Tao Y 2007 Luminescent materials and spectroscopic properties of Dy^{3+} ion *J. Lumin.* **122–123** 927–30
[42] Kaur S, Rao A S and Jayasimhadri M 2019 Color tunability and energy transfer studies of Dy^{3+} / Eu^{3+} co-doped calcium aluminozincate phosphor for lighting applications *Mater. Res. Bull.* **116** 79–88
[43] Kumari P and Manam J 2015 Structural, optical and special spectral changes of Dy^{3+} emissions in orthovanadates *RSC Adv.* **5** 107575–107584
[44] Dutta S, Som S and Sharma S K 2015 Excitation spectra and luminescence decay analysis of K^+ compensated Dy^{3+} doped $CaMoO_4$ phosphors *RSC Adv.* **5** 7380–7
[45] Sun J, Zhang X, Xia Z and Du H 2011 Synthesis and luminescence properties of novel $LiSrPO_4$:Dy^{3+} phosphor *Mater. Res. Bull.* **46** 2179–82
[46] Dahiya H, Dalal M, Singh A, Siwach A, Dahiya M, Nain S, Taxak V B, Khatkar S P and Kumar D 2020 Spectroscopic characteristics of Eu^{3+}-activated $Ca_9Y(PO_4)_7$ nanophosphors in Judd–Ofelt framework *Solid State Sci.* **108** 106341
[47] Ayoub I and Kumar V 2023 Synthesis, photoluminescence, Judd-Ofelt analysis, and thermal stability studies of Dy^{3+}-doped $BaLa_2ZnO_5$ phosphors for solid-state lighting applications *RSC Adv.* **13** 13423–37
[48] Phogat P, Taxak V B and Malik R K 2022 Crystallographic and optical characteristics of ultraviolet-stimulated Dy^{3+}-Doped $Ba_2GdV_3O_{11}$ Nanorods *J. Electron. Mater.* **51** 4541–54

2024 IEEE 9th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 5-7, 2024

Dynamic Resource-Aware Federated Framework for Secure and Sustainable Learning in Data Sensitive Applications

Manu Narula
Department of Information Technology
Delhi Technological University
Delhi, India
manunarula1996@gmail.com

Jasraj Meena
School of Computer and System Sciences
Jawaharlal Nehru University
Delhi, India
jasrajengg@gmail.com

Dinesh Kumar Vishwakarma
Department of Information Technology
Delhi Technological University
Delhi, India
dvishwakarma@gmail.com

2024 IEEE 9th International Conference for Convergence in Technology (I2CT) | 979-8-3503-9447-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/I2CT61223.2024.10543390

***Abstract*— Federated Learning (FL) quickly gained popularity as a secure distributed learning technology based on insights sharing. The client-level privacy provided by FL has enabled its extensive integration in fields dealing with susceptible data such as E-commerce, the Internet of Medical Things, etc. However, the transmitted insights may still be subjected to eavesdropping since the security of the transmission medium cannot be guaranteed. Using complex techniques, this can lead to the inference of the client data. In many cases, quantization is applied over transmitted payload to hinder such attempts but results in drop of accuracy in the global models. Moreover, the capabilities of client nodes are not considered in most implementations which may lead to under or over-utilization of client resources and cause client dropouts due to exhaustion. When combined, the quantization and client dropouts cause a significant performance loss and make the network less stable. In this paper, we propose a Dynamic Resource-Aware Federated Framework for Secure and Sustainable Learning, which layers the quantization process with an active workload management technique based on the capabilities of the client nodes. The framework minimizes the dropout rates significantly while recovering the accuracy lost in quantization to some extent. Through experiments compared against standard FL and Quantized FL over multiple datasets, the DRAFFSS shows gains in security over the former and in accuracy over the latter while keeping client dropouts negligible.**

Keywords—federated learning, data-sensitive applications, security, resource optimization, artificial intelligence

I. Introduction

Artificial Intelligence (AI) has made its way into every conceivable sphere of life to varying extents in recent years. Ranging from medical domain to automated heavy machinery AI has impacted execution, production, or planning phases section either directly or indirectly [1]. The underlying models in these applications often demand massive amounts of data in their training phase and in some cases, throughout their usage cycle for continuous tuning. In many instances, the required data is sensitive and can be used to infer crucial information about data owners leading to issues including cybercriminal activities. To cater to the growing concern regarding data privacy, Federated Learning (FL) was proposed which makes use of insights sharing in place of raw data [2] in a distributed node-based environment. The insights/gradients shared can then be used to modify the global results as required.

A. Federated Learning

FL had its inception in 2016 when McMahan et al. [2] proposed parameter transmission over a distributed learning network. The method provided a higher level of privacy compared to traditional AI methods and paved the way for collaborative learning efforts expanding both, the dataset variations and computing power. There exist many types and subtypes of FL. On a large scale, FL is categorized as Horizontal FL (HFL) [3] and Vertical FL [4]. In HFL, the training data available to different devices is identical in terms of feature set. In VFL, the data partitioning incorporates varied attributes throughout the subsets which are handled by a host controller. The correlation keys for entries that house data from different owners in separate columns in the same row are calculated by applying Private Set Intersection (PSI) [5]–[7].

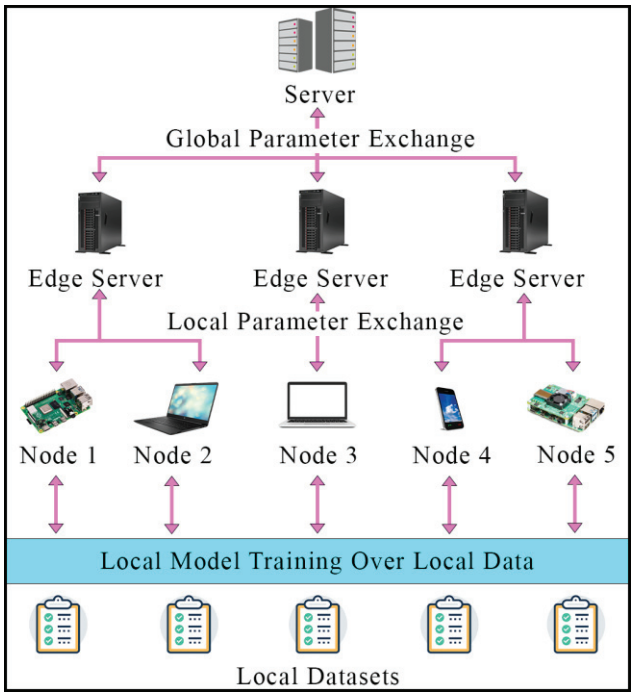


Fig. 1. Standard Federated Learning

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:39:59 UTC from IEEE Xplore. Restrictions apply.

Fig. 1. shows the basic architecture of an FL setup. Here multiple client nodes (mobile, laptop) communicate with a cloud/edge server to synchronize model parameters. These parameters are used to carry out localized training on client nodes. The updated set of parameters is uploaded back to the server for aggregation into the central model. The process is repeated over multiple rounds each time with a randomly selected subset of devices.

Incentive-Aware FL (IFL) [8]–[10] is a special sub-type of FL that tracks the contributions made by the participating nodes in the network. This record is used to allot rewards that may differ depending on the type of contribution or application. The associated reward system attracts new nodes to the network and refreshes dead or idle nodes to participate.

B. Data-Sensitive Applications

A Data-Sensitive Application (DSA) is an application that deals with data that is highly sensitive to the clients. These may include data corresponding to the social footprint [11], medical records [12], or economic status [13]. Most of these applications greatly increase the ease and quality of life. But, also makes private information more susceptible to eavesdropping. FL provides a level of immunity to DSAs from these issues.

However, many analytical methods have been developed to infer confidential data from transmitted parameters in FL setups [14]. One of the often-applied countermeasures is parameter quantization which incurs an accuracy bump proportional to the extent of quantization [15], [16] known as the quantization loss. This loss can be negated to some extent by stabilizing the network and reducing staggering nodes. Among the options available, the most promising optimizing method for less staggering is seen in the works of Li et al. [17] who implemented a modified version of Incentive-Aware FL as Workload-Aware FL. The proposed work adds a layer of server-side computation after each training round presenting a necessary but minimal cost. As evident by studies, the current state-of-the-art techniques focus either on security or efficiency but none tries to balance both. We propose a hybrid framework that employs the workload-aware FL over quantization in an FL setup to balance the quantization loss and ensure security if a gradient/parameter leak is encountered. We summarize the main contributions of this paper as follows:

- We propose a Dynamic Resource-Aware Federated Framework for Secure and Sustainable Learning (DRAFFSS) in DSAs.
- To better balance the quantization loss incurred, Workload-Aware FL is employed.
- Through extensive experimentation on various medical classification datasets, we show that our frame-work provides security at negligible accuracy loss and staggering nodes compared to existing techniques.

The rest of the paper is structured as follows: Section 2 outlines the related work done. Our proposed framework is discussed in section 3 and the experimental setup and results are elaborated in section 4. Finally, section 6 concludes the paper and provides potential future directions.

II. Related Work

New threats, solutions, and improvements emerge daily in technology fields, which make the state-of-the-art systems obsolete rapidly. FL is no stranger to this trend as the eavesdropping of the minimal transmitted parameters can allow attackers to invert the model and recreate the used dataset with perfect accuracy in some cases [14], [18]. Researchers have come forward with many potential solutions to prevent this inference. One of the most widely used techniques is quantization of parameters.

A. Quantization

There has been considerable work done in the field of quantization at certain levels. Initially, quantization was applied to the weights of a Neural Network (NN) model to optimize resources. Works of Li et al. [16] and Bao et al. [19] highlight attempts made toward weight-based model compressions. Choi et al. [20] proposed quantization for activation function, which compresses the activation functions during the training phase, while Li et al. [15] proposed an approach to over-come performance degradation in quantized Super Resolution NN. However, both these approaches are reliant on training, making them computationally expensive. Fang et al. [21] proposed a technique to estimate the quantized parameters and identify the optimal breakpoints to minimize the quantization loss. However, using quantization identically for parameters with dissimilar probability distributions results in limited success. Dong et al. [22] proposed a layer-wise quantization approach using the Hessian spectrum of layers for the selection of quantization precision optimized for each layer. Started as a method for resource optimizing, quantization showcases the potential for boosting security in FL applications.

FL natively improves the security and privacy of the involved data by limiting the data transfer across the network to model parameters. Some recent studies demonstrated the reconstruction of private data inferred from eavesdropping on these parameters. Xu et al. [23] proposed a quantization-based enhancement to put a check on the effectiveness of these inferences. Using 3-bit and 5-bit quantization with clipping over various datasets, the proposed model shows considerably less quantization loss compared to state-of-the-art Apot and Uniform quantizers. The approach however suffers from a loss in accuracy by 0.02% to 15% proportional to the complexity of the dataset used, restricting the utility toward DSAs.

Kerkouche et al. [24] presented another security approach partially focusing on resource optimization in low-powered networks like the Internet of Medical Things (IoMT). The results highlight significant drops in bandwidth consumption. This drop reduces the strain on resources in the network and minimizes the time taken per round of the FL algorithm. However, similar to the works of Xu et al., the experimentations show a loss in accuracy proportional to the complexity. IFL techniques over the base quantizer system can compensate for some of these performance gaps.

B. Incentive-Aware and Workload-Aware FL

IFL is a sub-type of FL that adds a step at the end of the training round to estimate the individual contributions of nodes to the global model. These contributions are evaluated to provide some form of incentive to the nodes to motivate further performance. Wang et al. [9] proposed an IFL

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:39:59 UTC from IEEE Xplore. Restrictions apply.

method by incorporating blockchain technology to induce trust and credibility between nodes. While proving to be a significantly stable, secure, and reliant technique, the overheads of blockchain [25] are difficult to negate in most use cases for resource-starved setups like IoMT. Nishio et al. [26] and Abdulrahman et al. [27] proposed client selection schemes using IFL. But both approaches proved to be either resource intensive or of low convergence due to hard constraints on participating nodes.

A modified implementation of IFL is proposed by Li et al. [17], which implements the contribution evaluation for clients based on their sustainable workloads. The approach enables the allocation of optimal workload in gradual steps to the nodes decreasing the chance of dropouts during training. We couple a modified version of the technique in [17] with a quantizing algorithm to form a hybrid algorithm that improves security and minimizes quantization loss and straggler nodes.

III. Proposed Framework (DRAFFSS)

This section describes in detail, our hybrid approach combining the security aspect of quantization with optimized throughput gained by virtue of workload-aware load selection. A detailed view of the complete frame-work with all the sub-functions is shown in Fig. 2.

At the start of each round, a subset of client nodes is randomly selected and the shortlisted nodes are queried about their capabilities. The client nodes respond with the lower & upper bound epoch, and the maximum sustainable epoch by the device. These nodes are provided with the global model parameters (step 1). The devices taking into account the individually calculated workload for the round, train the model locally. After training the updated models are quantized based on the quantization boundaries of the parameters. These boundaries are employed to serve two purposes; To provide security in case the transmitted parameters are compromised and to minimize the network load. A copy of the original model is stored on the client node and the quantized model is uploaded to the server for aggregation and updating the global model (step 6). The process may run for multiple rounds de-pending on the required level of training and the variety of data in the client nodes.

A. Security: Quantization

The security aspect of the approach relies on the quantization of model parameters while maintaining minimal quantization loss. An element in consideration can be quantized toward the lower or upper bound as feasible. The probabilities of an element x being quantized towards either boundary value is calculated as follows:

$$Pr_{x \to a_i - 1} = \frac{a_i - x}{a_i - a_{i-1}} \tag{1}$$

$$Pr_{x \to a_i} = \frac{x - a_i}{a_i - a_{i-1}} \tag{2}$$

Given that all parameters of the model are arranged in form a vector $\overrightarrow{p_v}$, having a probability distribution $f(x)$. For an element x falling between boundaries a_{i-1} and a_i and the total squared quantization error σ^2 is calculated as:

$$\sigma^2 = \sum_{i=0}^{2^b-1} \int_{a_{i-1}}^{a_i} [-x^2 + (a_{i-1} + a_i)x - a_i a_{i-1}] f(x) dx \tag{3}$$

Using partial derivatives, the authors in [23] derived quantization probability of ith element having minimum quantization loss as:

$$Pr_i = \frac{a_{i+1} \int_{a_{i-1}}^{a_{i+1}} f(x)dx - \int_{a_{i-1}}^{a_{i+1}} x f(x) dx}{a_{i+1} - a_{i-1}} \tag{4}$$

Algorithm 1 elaborates on the quantization employed in an FL setup. At the initiation of each training round 't', a subset of client nodes 'R' is selected by the server. All nodes in R receive the model from the global server. This model is denoted as m_0 or m_{t-1} (line 1). The clients then train the model k times over local data and update the model to m_t (lines 3-6). We then quantize the model by applying (4) and rounding off the parameters accordingly using stochastic rounding (line 7-8) to create $m_{\sim t}$. The final model m_t is kept with the client while quantized $m_{\sim t}$ is uploaded to the server for aggregation.

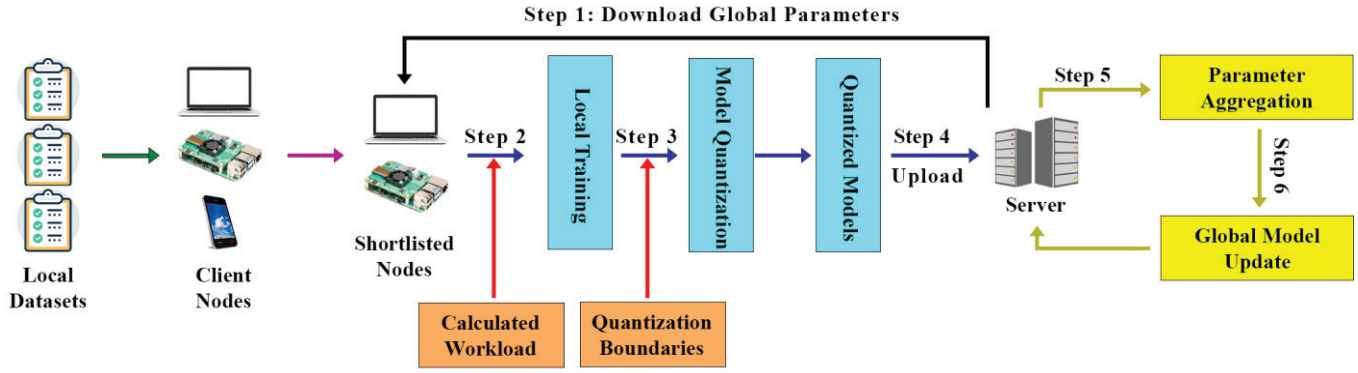


Fig. 2. Architecture of Proposed Framework (DRAFFSS)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:39:59 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Quantization of Parameters in FL Setup (Fed-q)

Input: Initial model m_0;
Output: Final model m_t;

1 **function Fed-q** (m_0)
2 $m_0 \leftarrow m_{t-1}$;
8 $m_{\sim t} \leftarrow Q\left(m_t, \{a_\tau^t\}_{j=0}^{2^b-1}\right)$; //stochastic rounding
9 Uploads $\{a_\tau^t\}_{j=0}^{2^b-1}$ and $f(m_{\sim t})$ to server;
10 **return** m_t;

B. Optimization:Client Selection

Quantization strengthens the security of the data from inference attacks, but the induced compression results in loss of accuracy in the performance of the global model. This restricts the usage in critical applications like the DSAs. To compensate for this loss, a workload optimization technique is integrated based on [17]. The authors propose workload-aware client selection and showed improvements in the accuracy of the trained model compared to traditional FL. The approach also highlights a stable network with minimal stragglers. We layer this approach over quantization to regain the accuracy lost in algorithm 1. The authors define the predicted workload for a client in a particular round by increasing the workload of the same client in the previous round by a fraction of itself. If the client drops out of the network, then the workload for that client in the next round is halved. Both cases are depicted by (5).

$$\hat{E}_k^{t+1} = \begin{cases} \frac{1}{2}\hat{E}_k^t, if\ client\ drops\ out \\ \hat{E}_k^t + \frac{\mu}{\hat{E}_k^t}, otherwise \end{cases} \quad (5)$$

Here $\hat{E}_k^t$ is the workload predicted for client k in round t, and μ is a hyper-parameter defined on an application-to-application basis. The selection of value of μ is crucial because if the value is too high, the stragglers will rise and if the value is too low, the convergence time will be infeasible. Algorithm 2 shows process for calculation of optimal workload for a client in the next round.

The algorithm takes input for the sustainable capacity of the client to process workload: $\tilde{E}_k^t$, the lower and upper limits of epoch for workload assignment: $\hat{L}_k^t$ and $\hat{H}_k^t$. Lines 2-5 initiates the second portion of (5). The first portion of (5) is handled by lines 6-10. Lines 11-14 are triggered when the lower limit exceeds the sustainable workload value for the client. It sets the workload of next round as NULL and halves both limits as per (5) highlighting a client drop from the algorithm.

Algorithm 2: Workload Prediction (Round_Load)

Input: Locally sustainable workload $\tilde{E}_k^t$ of client k at round t, the lower limit of the epoch for $\tilde{E}_k^t$, $\hat{L}_k^t$, the upper limit of the epoch for $\tilde{E}_k^t$, $\hat{H}_k^t$.
Output: the predicted minimum workload $\hat{L}_k^{t+1}$, the predicted maximum workload $\hat{H}_k^{t+1}$, the training workload $\hat{E}_k^t$.

1 **function Round_Load** ($\hat{L}_k^t, \hat{H}_k^t, \tilde{E}_k^t$)
2 **if** $\tilde{E}_k^t$ is greater than $\hat{H}_k^t$ **then**
3 $\hat{L}_k^{t+1}, \hat{H}_k^{t+1} \leftarrow \hat{L}_k^t + \frac{u}{\hat{L}_k^t}, \hat{H}_k^t + \frac{u}{\hat{H}_k^t}$
4 $\hat{E}_k^t \leftarrow \hat{H}_k^t$
5 **end**
6 **else if** the actually affordable workload falls into $[\hat{L}_k^t, \hat{H}_k^t]$ **then**
7 $\hat{L}_k^{t+1} \leftarrow min\left(\hat{L}_k^t + \frac{u}{\hat{L}_k^t} + \frac{1}{2}\hat{H}_k^t\right)$
8 $\hat{H}_k^{t+1} \leftarrow max\left(\hat{L}_k^t + \frac{u}{\hat{L}_k^t} + \frac{1}{2}\hat{H}_k^t\right)$
9 $\hat{E}_k^t \leftarrow \hat{L}_k^t$
10 **end**
11 **else**
12 $\hat{L}_k^{t+1}, \hat{H}_k^{t+1} \leftarrow \frac{1}{2}\hat{L}_k^t, \frac{1}{2}\hat{H}_k^t$
13 $\hat{E}_k^t \leftarrow 0$
14 **end**
15 **return** $\hat{L}_k^{t+1}, \hat{H}_k^{t+1}, \hat{E}_k^t$

C. Hybrid Algorithm

Combining the two aspects of security and optimization from the discussed works, we propose Dynamic Resource-Aware Federated Framework for Secure and Sustainable Learning to allow secure training with minimal straggler nodes and accuracy loss. The framework employs a 5-bit quantizer for parameter compression to optimize bandwidth consumption and prevent inference in case of eavesdropping on the network. The quantization loss induced is compensated by adding a layer for dynamic workload-allotment which reduces the chance of client drop out by allocating gradually increasing workload to the devices. The complete process of DRAFFSS is explained in Algorithm 3. At the start of each communication round a subset of clients is selected based on a probability list and global weights are shared to these clients (lines 1-3). The clients respond with the respective sustainable and minimum workloads (line 4). These values are passed on to a function employing algorithm 2 to determine workloads for the ongoing round which are passed to Fed-q function (algorithm 1) to carry out training (lines 5-7). Finally, the weights returned by Fed-q are aggregated into the global model (line 9).

Algorithm 3: DRAFFSS

Input: Number of Training Rounds T and list of clients in the network R
Output: Global trained model weights w^T

1 **for** each training round t = 1 to T **do**
2 $C_t \leftarrow$ shortlist clients based on probability list generated for R.
3 server passes the global weights w_t to shortlisted clients.
4 $\{(\hat{L}_k^t, \hat{H}_k^t, \tilde{E}_k^t)\} \leftarrow$ receive reply from C_t. // Request initial workloads and sustainable limits
5 **for** each client $k \in C_t$ in parallel **do**
6 $(\hat{L}_k^t, \hat{H}_k^t, \hat{E}_k^t) \leftarrow$ **Round_Load** $(\hat{L}_k^t, \hat{H}_k^t, \tilde{E}_k^t)$ //**Algorithm 2**
7 $\Delta_k^t, H_k^t \leftarrow$ **Fed-q**(m_k) //trained with workload values $(\hat{L}_k^t, \hat{H}_k^t, \hat{E}_k^t)$, //**Algorithm 1**
8 **end**

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:39:59 UTC from IEEE Xplore. Restrictions apply.

9 $w^{t+1} \leftarrow w^t - \frac{\sum_{k \in C^t} \Delta_k^t}{\sum_{k \in C^t} h_k^t}$

10 **end**

D. Baseline Model

We have used Resnet-18 [28], an 18-layer model with ReLU activation. We used SGD optimizer with adaptive learning rate of 0.1 dividing by 10 when error graph plateaus similar to that used in [14]. The exact model used is depicted in Fig. 3.

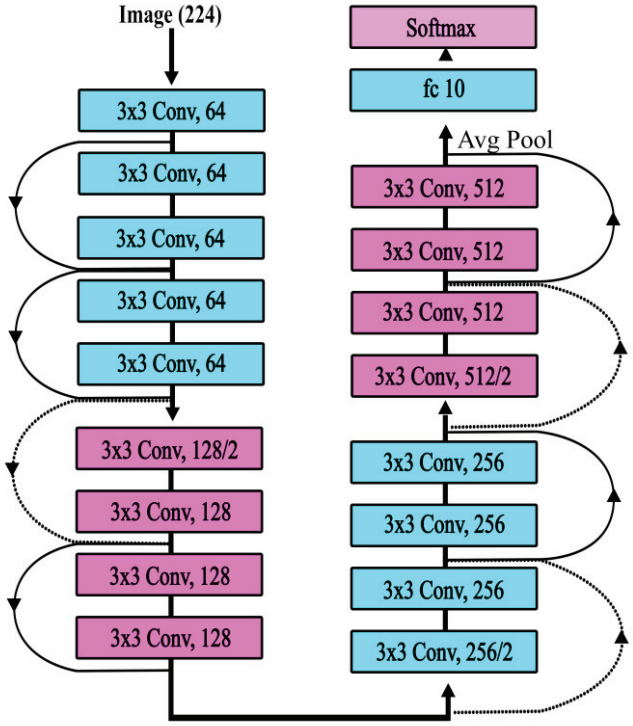


Fig. 3. Architecture of the baseline Resnet 18 model

IV. Results and Analysis

This section elaborates on our experimental environment and results over the various datasets used. We compared our model with standard FL as well as FL with 5-bit quantization to present a detailed view of the gains achieved by the hybrid approach.

A. Experimental Setup

Our simulations are carried out on a system running Windows 11 Pro version 22H2 equipped with an Intel(R) Xeon(R) Silver 4114 CPU @ 2.20GHz, 128GB DDR4 RAM, and two NVIDIA Titan RTX GPUs in SLI-Crossfire having 24GB VRAM each. We implemented our models in Python 3.9.16 with Cuda 12.1 [29], Tensorflow 2.12 [30] , and Pysyft 0.9.0 [31] over Jupyter Notebook in Conda virtual environment.

To replicate a network with heavy traffic and constraints implied by the bandwidth, we simulate 1000 nodes with slightly varying resources such as more or less RAM, CPU-cycles, and data-location (GPU/RAM).

B. Dataset Description

We tested our algorithm against the implementation discussed in [23] and standard FL over two publicly available medical image datasets having dissimilar complexity and sample size. A brief description of the datasets and their composition is as follows.

1) Diabetic Retinopathy

Diabetic Retinopathy Dataset [32] is a collection of high-resolution retinal images labelled 0 through 4 depicting absence or presence of diabetic retinopathy in the patient. There are two separate images for left and right eye for each patient. Fig. 4. Shows some sample images from the Diabetic Retinopathy Dataset.

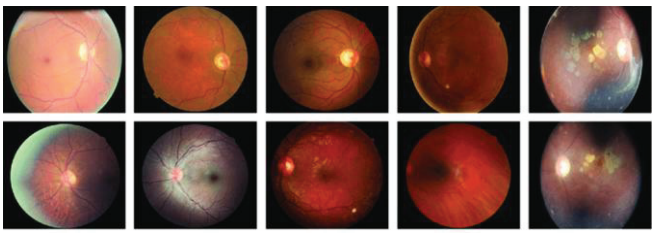


Fig. 4. Sample images from Diabetic Retinopathy Dataset

2) RSNA Bone Age

RSNA Bone Age Dataset [33] is an image dataset housing x-ray scans of hands of multiple children. The aim of the dataset is to predict both age and gender of the individuals by analyzing bone structures. Fig. 5. shows some sample images from the dataset.

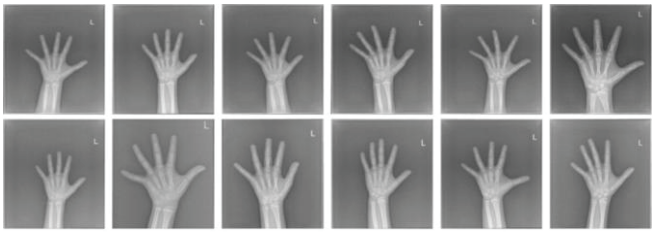


Fig. 5. Sample images from RSNA Bone Age Dataset

C. Results

We benchmarked DRAFFSS over the two datasets discussed in previous section and evaluated it against standard FL and Fed-MSQE[23] with the 5-bit quantization mode. Primarily four metrics are used for evaluation, Accuracy, Straggler node percentage, Precision, and Recall. These are defined as follows.

$$Accuracy = \frac{TrueNegatives(TN) + TruePositives(TP)}{TN + TP + FN + FP} \quad (5)$$

$$Straggler\% = \frac{\sum_1^T \frac{Nodes\ active\ in\ C_t * 100}{1000}}{T} \quad (6)$$

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

Our DRAFFSS demonstrated improved accuracy compared to the Fed-MSQE while also showing significantly less straggler nodes in all the comparisons. A summary of results attained by DRAFFSS is shown in table I.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:39:59 UTC from IEEE Xplore. Restrictions apply.

TABLE I. RESULTS ACHIEVED BY DRAFFSS IN VARIOUS DATASETS

Dataset	Accuracy	Straggler Node (%)
Diabetic Retinopathy	79.20	14
RSNA Bone age	82.60	12

a) Comparison in Diabetic Retinopathy Dataset.

Being the simpler of the two datasets, the results of all techniques in comparison appear close, well within the margin of error due to randomness. The standard FL tops the accuracy curve, Fed-MSQE shows loss of roughly 5.70%, and Fed-WALQ shows a loss of just 2.50% highlighting significant improvement in the narrow range available. A comparison of test accuracies of the three techniques over Diabetic Retinopathy Dataset is shown in Fig. 6.

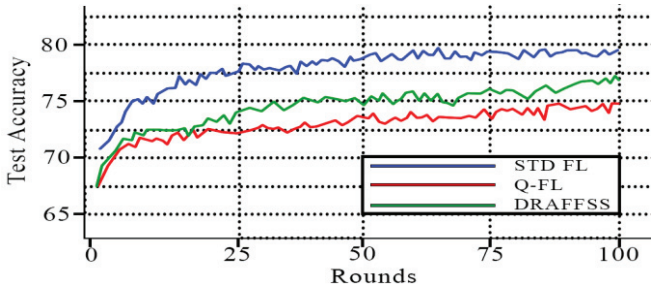


Fig. 6. Test Accuracy Comparison for Diabetic Retinopathy Dataset

Due to quantization and workload allotment, the number of stragglers is also seen to be reduced significantly when compared to traditional FL and Fed-MSQE as highlighted by the graph in Fig. 7.

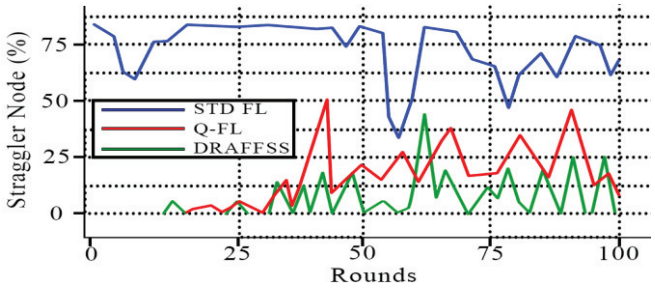


Fig. 7. Comparison of straggler nodes present during training over Diabetic Retinopathy Dataset

b) Comparison in RSNA Bone Age Dataset.

The dataset is designed to predict age of the patient based on x-rays. Hence the accuracy tested corresponds to the predictions that lie within 5.4 months of the actual ages. Here the standard FL attained a testing accuracy of 90%. While Fed-MSQE attained a testing accuracy of 77.2%, DRAFFSS attained an accuracy of 82.6% highlighting a drop of only 8.2% against 14.2 % of Fed-MSQE. The comparison of testing accuracy for all three techniques is shown in Fig. 8.

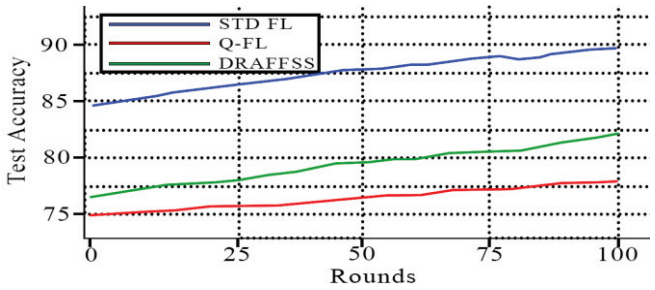


Fig. 8. Test Accuracy Comparison for RSNA Bone Age Dataset

Comparing the straggler percentages, standard FL suffered from 73.5% straggler node presence. While Fed-MSQE and DRAFFSS dealt with 14% and 12% presence respectively. Fig. 9. Shows the straggler node variations throughout training for the three techniques.

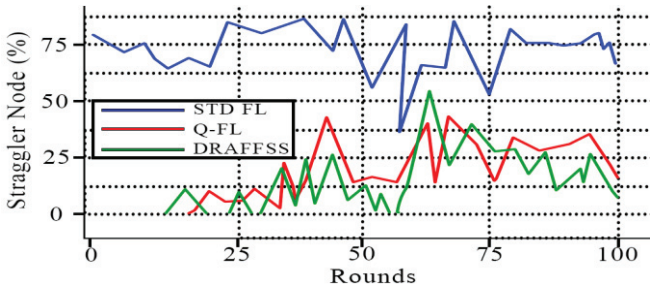


Fig. 9. Comparison of straggler nodes present during training over RSNA Bone Age Dataset

A summary of all the test accuracy results is tabulated in Table II.

TABLE II. ACCURACY AND STRAGGLER% ACHIEVED BY STANDARD FL, FED-MSQE, AND DRAFFSS OVER THE TWO MEDICAL DATASETS

Method\ Dataset	Diabetic Retinopathy		RSNA Bone Age	
	Test Acc	Straggler %	Test Acc.	Straggler %
Std. FL	79.80	72.00	90.00	73.50
FED-MSQE (5-bit)	75.00	17.00	77.20	14.00
DRAFFSS	77.50	14.00	82.60	12.00

TABLE III. PRECISION AND RECALL ACHIEVED BY STANDARD FL, FED-MSQE, AND DRAFFSS OVER THE TWO MEDICAL DATASETS

Method\ Dataset	Diabetic Retinopathy		RSNA Bone Age	
	Precision	Recall	Precision	Recall
Std. FL	79.76	79.79	90.01	90.00
FED-MSQE (5-bit)	75.12	75.12	77.20	77.21
DRAFFSS	77.50	77.49	82.60	82.60

Based on the results observed, DRAFFSS demonstrates its capability to compensate for performance loss induced by quantization. The compensation maybe marginal or significant depending on the complexity of the data used for training. The reduction in straggler nodes however is significant in all the test cases highlighting a stable system compared to the existing techniques. The improved stability allows the network to train more effectively as there are few interruptions due to node dropouts.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:39:59 UTC from IEEE Xplore. Restrictions apply.

Accessing from the point of view of the node devices, the workload allotment helps keep resources from being over utilized. This in turn enables the devices to continue their local routines without issues.

V. Conclusion and Future Scope

A. Conclusion

In this paper, we propose Dynamic Resource-Aware Federated Framework for Secure and Sustainable Learning (DRAFFSS), a hybrid FL framework for DSAs, aimed at hindering inference of private data in case of an eavesdropping attack. In addition, the quantization reduces the transmission load optimizing the bandwidth use for constrained networks like IoMT. The framework uses a state-of-the-art 5-bit quantization scheme (Fed-MSQE) and layers it with adaptive work-load allotment scheme. The client devices are allotted sustainable workloads adapting to the work history of the client device reducing the dropout chances in the network and boosting throughput during training rounds. Experimental results across two medical datasets show significant improvements in straggler node ratio and testing accuracy over pure quantization methods ensuring a substantial level of security.

B. Future Scope

The improvements achieved in accuracy during our simulations are highly dependent on the complexity associated with test datasets. Development of a generic framework less affected by complexities of data is the most prominent direction for future research.

References

[1] I. Ahmed, G. Jeon, and F. Piccialli, "From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where," *IEEE Trans Industr Inform*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022, doi: 10.1109/TII.2022.3146552.

[2] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Feb. 2017.

[3] D. Liu, L. Bai, T. Yu, and A. Zhang, "Towards Method of Horizontal Federated Learning: A Survey," in *2022 8th International Conference on Big Data and Information Analytics (BigDIA)*, IEEE, Aug. 2022, pp. 259–266. doi: 10.1109/BigDIA56350.2022.9874186.

[4] K. Wei *et al.*, "Vertical Federated Learning: Challenges, Methodologies and Experiments," Feb. 2022, [Online]. Available: http://arxiv.org/abs/2202.04309

[5] K. Nomura, Y. Shiraishi, M. Mohri, and M. Morii, "Secure Association Rule Mining on Vertically Partitioned Data Using Private-Set Intersection," *IEEE Access*, vol. 8, pp. 144458–144467, 2020, doi: 10.1109/ACCESS.2020.3014330.

[6] A. Abadi, S. Terzis, R. Metere, and C. Dong, "Efficient Delegated Private Set Intersection on Outsourced Private Datasets," *IEEE Trans Dependable Secure Comput*, vol. 16, no. 4, pp. 608–624, Jul. 2019, doi: 10.1109/TDSC.2017.2708710.

[7] Z. Wang, K. Banawan, and S. Ulukus, "Private Set Intersection Using Multi-Message Symmetric Private Information Retrieval," in *IEEE International Symposium on Information Theory - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Jun. 2020, pp. 1035–1040. doi: 10.1109/ISIT44484.2020.9174221.

[8] Z. Wang, B. Yan, and A. Dong, "Blockchain Empowered Federated Learning for Data Sharing Incentive Mechanism," in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 348–353. doi: 10.1016/j.procs.2022.04.047.

[9] Z. Wang, Q. Hu, R. Li, M. Xu, and Z. Xiong, "Incentive Mechanism Design for Joint Resource Allocation in Blockchain-based Federated Learning," *IEEE Transactions on Parallel and Distributed Systems*, May 2023, doi: 10.1109/TPDS.2023.3253604.

[10] Y. Zhu, Z. Liu, P. Wang, and C. Du, "A dynamic incentive and reputation mechanism for energy-efficient federated learning in 6g," *Digital Communications and Networks*, Apr. 2022, doi: 10.1016/j.dcan.2022.04.005.

[11] H. Lin, K. Kaur, X. Wang, G. Kaddoum, J. Hu, and M. M. Hassan, "Privacy-Aware Access Control in IoT-enabled Healthcare: A Federated Deep Learning Approach," *IEEE Internet Things J*, 2021, doi: 10.1109/JIOT.2021.3112686.

[12] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nat Mach Intell*, vol. 2, no. 6, pp. 305–311, Jun. 2020, doi: 10.1038/s42256-020-0186-1.

[13] M. Schreyer, T. Sattarov, and D. Borth, "Federated and Privacy-Preserving Learning of Accounting Data in Financial Statement Audits," in *3rd ACM International Conference on AI in Finance*, New York, NY, USA: ACM, Nov. 2022, pp. 105–113. doi: 10.1145/3533271.3561674.

[14] Z. Li, L. Wang, G. Chen, Z. Zhang, M. Shafiq, and Z. Gu, "E2EGI: End-to-End Gradient Inversion in Federated Learning," *IEEE J Biomed Health Inform*, pp. 1–13, 2022, doi: 10.1109/JBHI.2022.3204455.

[15] H. Li *et al.*, "PAMS: Quantized Super-Resolution via Parameterized Max Scale," in *Lecture Notes in Computer Science*, Springer, Cham, 2020, pp. 564–580. doi: 10.1007/978-3-030-58595-2_34.

[16] Z. Li, J. Chen, L. Wang, B. Cheng, J. Yu, and S. Jiang, "CNN Weight Parameter Quantization Method for FPGA," in *2020 IEEE 6th International Conference on Computer and Communications, ICCC 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 1548–1553. doi: 10.1109/ICCC51575.2020.9345248.

[17] L. Li *et al.*, "Federated learning with workload-aware client scheduling in heterogeneous systems," *Neural Networks*, vol. 154, pp. 560–573, Oct. 2022, doi: 10.1016/j.neunet.2022.07.030.

[18] G. Zhang, B. Liu, T. Zhu, M. Ding, and W. Zhou, "Label-Only Membership Inference Attacks and Defenses In Semantic Segmentation Models," *IEEE Trans Dependable Secure Comput*, 2022, doi: 10.1109/TDSC.2022.3154029.

[19] Z. Bao, K. Zhan, W. Zhang, and J. Guo, "LSFQ: A Low Precision Full Integer Quantization for High-Performance FPGA-Based CNN Acceleration," in *IEEE Symposium in Low-Power and High-Speed Chips, COOL CHIPS 2021 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Apr. 2021. doi: 10.1109/COOLCHIPS52128.2021.9410327.

[20] J. Choi, Z. Wang, S. Venkataramani, P. I.-J. Chuang, V. Srinivasan, and K. Gopalakrishnan, "PACT: Parameterized Clipping Activation for Quantized Neural Networks," May 2018, [Online]. Available: http://arxiv.org/abs/1805.06085

[21] J. Fang, A. Shafiee, H. Abdel-Aziz, D. Thorsley, G. Georgiadis, and J. Hassoun, "Post-Training Piecewise Linear Quantization for Deep Neural Networks," in *Lecture Notes in Computer Science*, Cham: Springer, 2020. doi: https://doi.org/10.1007/978-3-030-58536-5_5.

[22] Z. Dong, Z. Yao, A. Gholami, M. Mahoney, and K. Keutzer, "HAWQ: Hessian AWare Quantization of Neural Networks with Mixed-Precision," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1905.03696

[23] Z. Xu, Y. Guo, C. Chakraborty, Q. Hua, S. Chen, and K. Yu, "A Simple Federated Learning-based Scheme for Security Enhancement over Internet of Medical Things," *IEEE J Biomed Health Inform*, 2022, doi: 10.1109/JBHI.2022.3187471.

[24] R. Kerkouche, G. Ács, C. Castelluccia, and P. Genevès, "Privacy-preserving and bandwidth-efficient federated learning: An application to in-hospital mortality prediction," in *ACM CHIL 2021 - Proceedings of the 2021 ACM Conference on Health, Inference, and Learning*, Association for Computing Machinery, Inc, Apr. 2021, pp. 25–35. doi: 10.1145/3450439.3451859.

[25] Y. Qu, M. P. Uddin, C. Gan, Y. Xiang, L. Gao, and J. Yearwood, "Blockchain-Enabled Federated Learning: A Survey," *ACM Comput Surv*, vol. 55, no. 4, Mar. 2022, doi: 10.1145/3524104.

[26] T. Nishio and R. Yonetani, "Client Selection for Federated Learning with Heterogeneous Resources in Mobile Edge," Apr. 2018, doi: 10.1109/ICC.2019.8761315.

[27] S. Abdulrahman, H. Tout, A. Mourad, and C. Talhi, "FedMCCS: Multicriteria Client Selection Model for Optimal IoT Federated Learning," *IEEE Internet Things J*, vol. 8, no. 6, pp. 4723–4735, Mar. 2021, doi: 10.1109/JIOT.2020.3028742.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:39:59 UTC from IEEE Xplore. Restrictions apply.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[29] R. S. Dehal, C. Munjal, A. A. Ansari, and A. S. Kushwaha, "GPU Computing Revolution: CUDA," in *Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018*, Institute of Electrical and Electronics Engineers Inc., Oct. 2018, pp. 197–201. doi: 10.1109/ICACCCN.2018.8748495.

[30] M. Abadi *et al.*, "TensorFlow: A System for Large-Scale Machine Learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, Savannah, GA, USA: USENIX Association, Nov. 2016, pp. 265–283. Accessed: Jul. 23, 2023. [Online]. Available: https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi

[31] A. Ziller *et al.*, "PySyft: A Library for Easy Federated Learning," in *Federated Learning Systems*, Springer, Cham, 2021, pp. 111–139. doi: 10.1007/978-3-030-70604-3_5.

[32] "Diabetic Retinopathy Detection | Kaggle." https://www.kaggle.com/c/diabetic-retinopathy-detection (accessed Dec. 06, 2022).

[33] "RSNA Bone Age | Kaggle." https://www.kaggle.com/datasets/kmader/rsna-bone-age (accessed Dec. 06, 2022).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:39:59 UTC from IEEE Xplore. Restrictions apply.

Dynamics of Phase Transition in Quark-Gluon Plasma Droplet Formation under Magnetic Field

Agam K. Jha
Department of Physics, Kirori Mal college
University of Delhi,
Delhi-110007, India

Aviral Srivastava
Delhi Technological University,
Shahbad Daulatpur, Main Bawana Road,
Delhi - 110042, India

arXiv:2406.14172v1 [hep-ph] 20 Jun 2024

Abstract: Pre-existing density of states for a Quark-Gluon Phase, based on Thomas-Fermi and Bethe mode, is expanded by incorporation of new variables. Results from recent study indicate that perturbations in the form of a finite non-zero chemical potential μ, $\mathbb{B}$, dynamic thermal masses $\mathbb{M}$ and of course Temperature $\mathbb{T}$ are indeed vital to fully comprehend the formation and dynamics of QGP. Simulations depict an overall increase in the stability of QGP in the paradigm of the statistical model. On the top of Free Energy, Entropy and heat capacity are calculated for the phase transition. The overall qualitative behavior, of entropy or Heat Capacity determines the order of phase transition of the QGP. Investigation of order of phase transition is carried out in this study through Monte-Carlo based differential element, which ensures the inclusion of the randomness of the collisions at the particle colliders.

I. INTRODUCTION

Quantum Chromodynamics (QCD) outlays the varying strength of the strong interaction through the QCD Phase Diagrams. Temperature $\mathbb{T}$ and chemical potential μ determine the state of in which hadrons exist. QCD is formulized on the SU(3) group renders the calculations much complex. Incorporating new factors in an already complicated system exorbitantly increases computation time. Moreover, the sheer number of dimensions in the QCD Lagrangian and hence the action integral, suggests a statistical approach would be quite precise. Hence, this study extends the existing statistical model[1], consistent with URHIC experiments at the Brookhaven National Laboratory (BNL) and CERN. The current prevailing statistical model incorporates temperature and chemical potential[2, 3] as the primary thermodynamic variables, in accordance with the QCD phase diagram. The model successfully predicts the order of phase transition from a confined state to a deconfined state. That being first order, which essentially translates to a discontinuity in the entropy S against temperature (T). Recent experiments at CERN and BNL suggest the precise description of QGP evolution necessitates the incorporation of new thermodynamic variables, i.e. Magnetic Field. Heavy Au-Au/Pb-Pb ion collisions, need not be head-to-head collisions.

Hence it is plausible that the magnetic field is playing a critical role in the formation of the QGP and possibly its physical characteristics. It is important to note that the magnetic field considered here is the external magnetic field [4, 5] present in the environment wherein the QGP droplets form. Another motivation for the incorporation of magnetic fields arises from cosmological and astronomical observations. Primordial magnetic fields of the universe have governed the dynamics of the large-scale magnetic fields that are found in galaxies today. More importantly these primordial were also associated with the phase transitions and inflation in the early universe [6]. Since QGP phase transitions are a manifestation QCD interactions, magnetic field [7] must have an effect on the evolution, which interacts through colour charge. Since QGP prevailed in the very early phase of the universe, it is thus crucial to consider this aspect.

With an aim to foster a robust and concise model, simple additions have been made to the existing statistical picture. Temperature is a measure of the energy of the system. The collisions which lead to the formation of QGP droplets are relativistic. Further, again from cosmological constraints, the cosmological environment enabled relativistic regimes. Thus, instead of rest masses, dynamical relativistic masses have been used [8]. Dynamical masses are explicit functions of temperatures. The highlight of this magnetic field study is the parametrization of magnetic

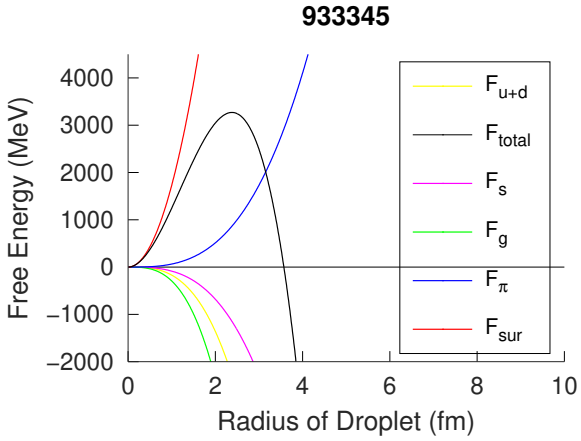


FIG. 1: Variation of Free Energy parametrized by T, μ and B. T = Tc = 175 MeV, μ=400 Mev, $eB = 15m_\pi^2$

field, as an indigenous attempt , strength [9] on the medium and all the constituents contributing to free energy, which in this case is the π-ionic medium.

II. DISCUSSION

A. Free Energy

Through the Thomas-Fermi Model of the atom and the Bethe Model for nucleons, the density of states for quarks and gluons is retained from the previous study cite[1]:

$$\rho_{q,g}(k) = \nu(1/\pi)^2 \left[[-V_{conf}(k)]^2(\frac{-dV_{conf}}{dk})\right] \tag{1}$$

Here V_{conf} is the confining potential [1, 10] for quarks and gluons, given by

$$[V_{conf}]_{q,g} = \frac{1}{2k}(\gamma_{q,g})g^2(k).T^2 - \frac{m_o^2}{2k} \tag{2}$$

All the following parameters are retained:

$$g^2(k) = (4/3)(12\pi/27)\left[\frac{1}{ln(1+k^2/\Lambda^2)}\right] \tag{3}$$

$$k_{min} = (\gamma_{q,g}.N.T^2.\Lambda^2/2)^{1/4} where N = (4/3)(12\pi/27) \tag{4}$$

$[-V_{conf}]_{q,g}$, the confining potential constructed from QCD constraints is retained, due to its apt description of the dynamical nature of quarks. While $\gamma_{q,g}$ are the hydrodynamic flow parameters, which account for the fluid nature of the plasma. Though a minor altercation is done with the original form of the potential, which is the addition of dynamic temperature-dependent masses. In this statistical model, the Peshier, and thus confining potential is invariant. Thus, the perturbative additions of magnetic field do not affect the Peshier potential constructed from Thomas-Fermi and Bethe models. There exists a simple explanation for this striking result. Magnetic field does no actual work, in the confinement of deconfinement processes, on the system. Peshier potential is a confining potential for the quarks and gluons. The thermal dynamic masses

are defined as [8]

$$\mathbb{M}_{d_{q,g}} = \gamma_{q,g}.T^2\left[\frac{1}{log(1+T/T_c)}\right]^2\left(\frac{16\pi^2}{11-\frac{2}{3}N_{f_{q,g}}}\right) \tag{5}$$

where T_c is the transition temperature, while $N_{f_{q,g}}$ is the flavour number which is taken to be 0 for gluons (bosons) and three flavours for quarks (up, down and strange).[8].

Thus the thermal mass modifies the Peshier potential to yield the following expression

$$[V_{conf}]_{q,g} = \frac{1}{2k}(\gamma_{q,g})g^2(k).T^2 - \frac{\mathbb{M}_{d_{q,g}}}{2k} \tag{6}$$

Magnetic field's interplay in the confining potential is disregarded in the context of this simple statistical model. Its interaction with the quark-gluon matter is reflected through the coupling in the partition function and hence ultimately the free energy of the plasma, with a plausible mediation by colour charge.

Through these quantities mentioned above, density of states and Peshier potential, the free energy of the QGP constituents is constructed which incorporates three thermodynamical variables as follows: -

$$F_i = -T.g_g\int dk.\rho_q(k).ln\left(1+e^{-\frac{\left(\sqrt{\mathbb{M}_q+k^2-(\mathbb{B})}-\mu_i\right)}{T}}\right) \tag{7}$$

F_i denotes the free energy for the constituents quarks of the QGP. Here we consider three flavors of quarks [u,d,s]: up, down and strange. Hence index i goes from 1-3. A similar modified construction can be defined for gluons.

$$F_g = T.g_g\int dk.\rho_g(k).ln\left(1-e^{-\frac{\left(\sqrt{\mathbb{M}_g+k^2-(\mathbb{B})}-\mu_g\right)}{T}}\right) \tag{8}$$

An additional negative is needed to make the gluon-ic free energy negative.$\pi - ionic$ medium also exists within the framework of the QGP formation. $\pi - ions$ manifest as a signature in detection of QGP [11] at particle colliders. It is highlighted that in this study the medium follows from the FD and BE statistics, as evident from the free energy integrand. Further, unlike the hadrons themselves, the π meson is not modelled to have the dynamic mass. It carries a fix mass of $\mathcal{M}_\pi \sim 138MeV$ [12].A strong reason for this characteristic is the proximal relation of π-ions to chiral symmetry restoration. Their sensitivity to QGP state variables like Temperature $\mathbb{T}$ and chemical potential $\mathbb{T}$ and magnetic field $\mathbb{B}$ holds vital importance under the context of this study.

The Free Energy Integral for the π-ionic medium is given by,

$$F_\pi = -\frac{3T\nu}{2\pi^2}\int dk.k^2.ln\left(1+e^{-\frac{\left(\sqrt{\mathcal{M}_\pi{}^2+k^2+(\mathbb{B})}-\mu_i\right)}{T}}\right) \tag{9}$$

The Free energy integrals maintain the consistency, which follows from the earlier study. Hence, again statistical distribution functions for Fermi-Dirac (quarks) and Bose-Einstein (Gluons) are being followed. Before addition of a magnetic field, the Free Energy integral follows a positive sign for quarks and a negative sign for gluons. Magnetic Field changes the nature of the integrals and reverses the trend. So, the algebraic signs preceding the integral flip, such an exercise is done on an ad-hoc basis and is vital to ensure the Free energy variation with the QGP droplet radius follows the path towards equilibrium.

Finally, the modified surface energy of the droplets is constructed in the presence of magnetic fields.

$$F_{sur} = \frac{\sqrt{2}R^2}{4}\left[\gamma_f T^3 + B * T\right] \tag{10}$$

In the above equation 10, $\gamma_f = \sqrt{(\frac{1}{\gamma_q})^2 + (\frac{1}{\gamma_g})^2}$ Dimension of magnetic field is MeV^2 in the natural units system. Coupling of the magnetic field with the system is dependent on the constituents. While for quarks, *pi* – ionic medium and surface energy, the field interacts as a square of a coupled quantity $\mathbb{B}$ and $\mathbb{T}$. On the other hand for bosons in the system, i.e. the gluons, the field interacts without a scaled-down factor of the temperature. Suggesting that coupling between fermions and hadrons is stronger as compared to gluons.

Limits in accordance with LHC	Magnetic Field Strength ($\frac{\times m_\pi^2 * 15}{0.303}$)
Minimum	0.98
	1.107
	1.2
	1.46
	1.71
	1.83
	2.00
Maximum	5.00

The range of magnetic field is fixed by the LHC exeperiments [13], while the supremum is imposed by numerical calculations done within this paradigm. π-ionic free energy and gluon-ic free energy are the the most susceptible to change in magnetic, with the former sets a much more substantial constraint.
The total free energy is the summation of all the individual constituents without any considerable interaction among them.

$$F_total = \left(\sum_{i=1}^{3} F_i\right) + F_g + F_\pi + F_{sur} \tag{11}$$

The existing free energies of the *pi* ionic medium and the surface free energy are parametrized with a magnetic field over the existing thermal dependence. Chemical potential does not play a part in the surface energy. It is modelled in a way which approximates the surface of the droplet in the phase space, which means momentum (k) which is ultimately due to temperature. Implying, that the surface energy (per unit area) is calculated for a single droplet is same for any droplet of some arbitrary radius and thus does not reflect a summation of the constituents of the system. Their dependence on the magnetic field is modelled which verifies the expected results. In the presence of large magnetic fields produced during the collisions at BNL and CERN, the QGP formation is significantly changed. Magnetic field, restricts the momentum space thus allowing for a greater density of Baryons and thus a higher energy density. Surface Free energy aptly appears like a parallelism to the surface tension of the Quark Gluon Plasma droplet: whose formation is governed by an interplay of QCD interactions and the three dynamical thermodynamic variables in this model. Chemical potential does not appear in the expressions, as explained above. The magnetic field governs the surface structure of each droplet formed through the deconfinement process.

Since the droplet is charged (QCD colour charge) and consists of charged constituents (quarks), the magnetic field affects its curvature [14]. At the equilibrium of the interplay between QCD forces, thermal energy and magnetic field, the droplet assumes a stable radius. *pi* – ionic free energy.

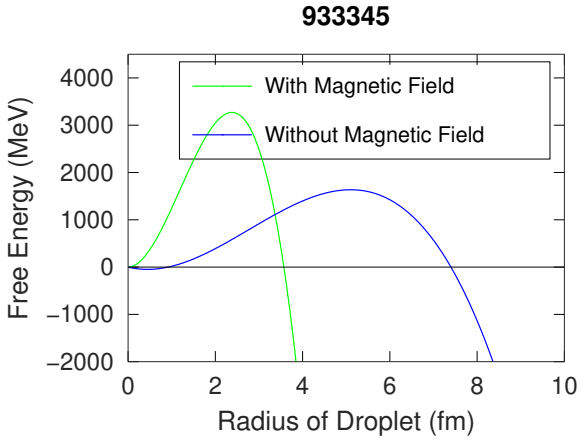


FIG. 2: Comparisson of stability in the absence (blue) and presence (green) of B. With a higher peak indicating a much more stable droplet. T = Tc , μ=400 Mev, $eB = 15m_\pi^2$

Magnetic field does not seem to affect the Peshier potential of the quarks and gluons. Magnetic field is an independent variable. Thus, making the macroscopic states of the fireball system as:
i) Temperature
ii) Chemical Potential
iii) Magnetic Field

B. Entropy, Heat Capacity and Phase Transition

An intricate perspective towards the formation of the de-confined phase of quarks and gluons is presented through the improved statistical model. To render a wholesome perspective though, a complete statistical and thermodynamic set of state functions is needed. Hence, through the free energy integrals, its derivative are calculated. These derivatives represent different thermodynamic state functions. First derivative of the free energy with temperature is computed, this quantity is the entropy (S) of the system. Subsequently, another derivative with temperature yield the heat capacity C_v .

As the system is transitioning towards equilibrium, entropy is expected to rise with temperature. Which means it will have a positive slope, which is nothing but the graph of the heat capacity. Further, the ratio of Entropy (S) and the heat capacity C_v define the speed of sound in the plasma phase. Speed of sound is accurately predicted by the QCD calculations. Hence, this quantity serves as a vital test for this extended simple statistical model.

$$\begin{aligned} S_{entropy} &= -\frac{\partial F}{\partial T} \\ C_{heat\ capacity} &= T\frac{\partial S}{\partial T} = -T\frac{\partial^2 F}{\partial T^2} \end{aligned} \tag{12}$$

Hence, the speed of sound can be defined as,

$$C_s^2 = S_{entropy}/C_{heat\ capacity} \tag{13}$$

In equation 13, C_s^2 is the square of the speed of sound. As per lattice calculations form Quantum Chromodynamics, this quantity has an supremum, capped at 1/3.

Entropy and heat capacity also yield information about phase transition. Since formation of Quark-Gluon Plasma involves a change of phase from a confined state of hadrons to a free or de-confined state, the process occurs through a thermodynamic phase transition. In the

previous study, with temperature as the only parameter, the order was phase transition was first order. Therefore, a discontinuity in the first derivative of free energy, i.e. entropy was observed. Verification of the nature of phase transition is done again. Mainly to check if its behavior is preserved or undergoes a change. Hence, the plots for entropy, heat capacity and speed of sound are studied to find the order.

III. RESEARCH METHODOLOGY

A. Literature Review

The study is reinstated by an overview of the existing statistical model. Followed by a thorough review of new research that has emerged in the field. ALICE experiments at CERN and the RHIC experiments at Brookhaven National Laboratory have fueled the research around this discipline. While the experiments take a long time to run, and an even longer time to collect, analyse annd study the results. Theoretical models, which are developed independently, aid in predicting the result of the experiments. Another vital role played by theoretical models, is in the process of defining constraints. Constraints allow a single run to be effective by dedicating the resources in a dedicated direction which is expected to yield results. Since the experiments reveal new information every time, the review procedure familiarizes with the existing sphere of knowledge around research in QGP. This study presents a novel and indigenous process to comprehend the complex dynamics of QGP formation, so the literature review provided the neccesary outlook which helped in scrutinising existing models and formulating the one employed here.

B. Data for Simulation

Computations form an integral part in the analysis, first mathematical model was formulated. Which was followed by a simulation through numerical and symbolic functions based computation. Hence, much of the data was generated within the framework of the simulation. As a result, there was no immediate need to collect data from previous studies. Though a variety of existing plots pertaining to the paradigm of the study were referred throughout the study.

C. Qualitative and Quantitative Analysis of results

Review process prepared a ground work to be built upon. A brief idea of the variation of Free energy in the position space was obtained. However, with the incorporation of every new variable, the system portrayed a different behaviour. The study concludes with an analysis of stability, through a comparison with previous models, analysis of order of phase transition and estimation of speed of sound near the transition temperature.

IV. RESULTS AND CONCLUSION

In absence of B thermal motion on account of alignment is small. As $\mathbb{B}$ enters the scenario, it competes with thermal energy and counters. So, one expects the degrees of freedom to change, hence the curve shifts to smaller radius values, which is the case. Indeed, this reduction in radius is observed only when the droplet forms i.e. free energy assumes a bell-like behaviour. The quantification of this decrease (in radius) should be proportional to the strength of the field. $\mathbb{B}$ thus has an apparent effect on the height of the peak which is proportional to the free energy of the system.

As a shift towards dynamic temperature-dependent masses is implemented. It is observed that

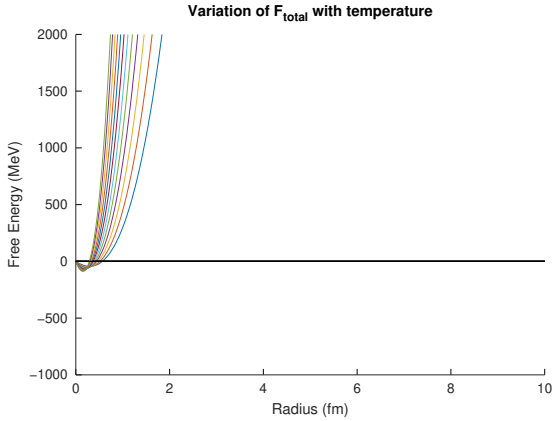


FIG. 3: Variation of free energy for different temperatures with $\mathbb{B} = 0$ and $\mu = 400$, with the ratio of flow parameters as 8.

droplets are not stable (monotonically increasing trends) until chemical potential is added to the distribution function. Which suggests dynamical masses increase the energy content of the system.

To reinstate the bell-shaped trends, chemical potential at first and magnetic field reduces the Free energy and enable formation of droplets with a finite peak. A new observation is $\mathbb{B}$ favours formation of smaller droplets, and hence brings down the mean stable radius of droplets. A high chemical potential is favored by the system ($\mu_q = \mu_g = 400\, MeV$). The constraint is set by the speed of sound, which is highly sensitive of the chemical potential. A numerical comparison reveals that at $T = T_c$, average over entire range of magnetic field for a radius of 4 fm shows a 16 times increase in the free energy peak as compared to the basic simple statistical model.

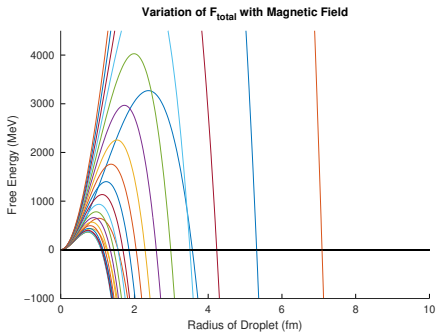
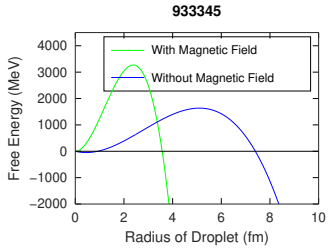
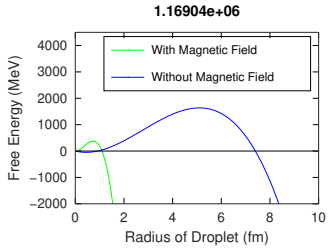
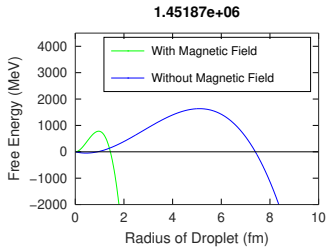
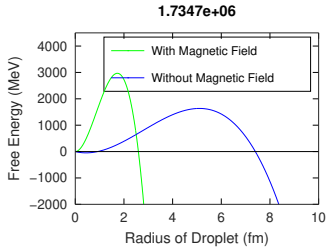


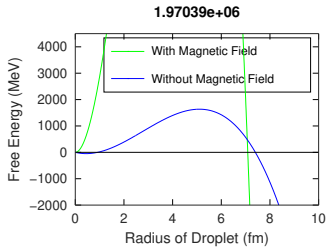
FIG. 4: Variation of Free Energy for various values of B with fixed T and μ











Dependence on the flow parameters has been reduced following the addition of temperature-dependent dynamical masses, chemical potential and magnetic field. The greatest factor affecting this change is chemical potential. Previously, altercation of the flow parameter ratios between quarks and gluons by a difference of 2, the stability of the plasma immensely shifted the trends of the Free Energy curves. The scenario has changed now. Although, the dependence has not vanished entirely, and a factor of $\gamma_g = (6, 8)\gamma_q$, still give a favorable peak, other variables simply assert dominance. Which means if Magnetic field is high enough, within the appropriate bounds, a feasible phase transition is observed outside the previous stable domain of flow parameters.

Finally, the order of Phase transition when all the thermodynamic variables are taken into account is a weakly first-order phase transition. There appears a distinct discontinuity in the trends of entropy near the transition temperature. However, a slight yet continuous deviation in

the case of heat capacity is also observed. The ratio of entropy and heat capacity gives the speed of sound (C_s^2). Comparing from the QCD results, which limits the C_s^2 at 1/3. Under the paradigm of this extended study, it is found that this natural bound is being respected by the trends. Further, the trends obey QCD calculations and computations based on ALICE data. At the effective temperature (222 MeV) [15], $C_{sT_{eff}}^2$ is within 16% margin of error in accordance with the bounds.

Concluding this study, it is emphasized that the stability of QGP is greatly enhanced. There are some new observations: dependency on flow parameters and reduced average stable radius which require further study.

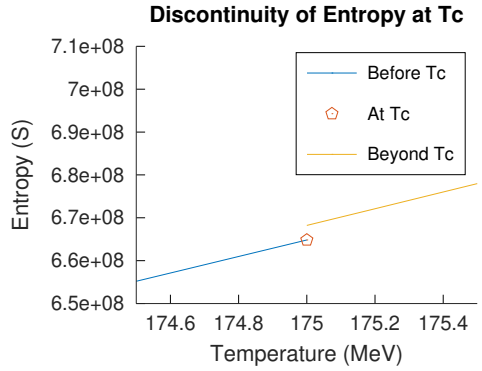


FIG. 5: Discontinuity of Entropy at Transition Temperature

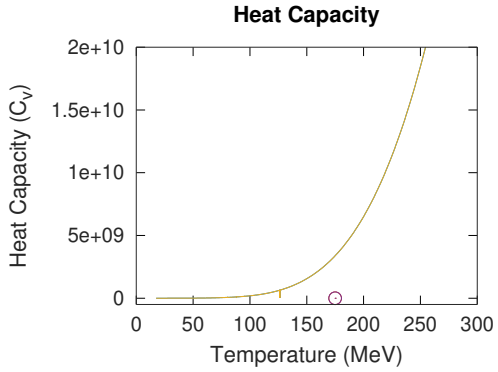


FIG. 6: Continuity of Heat Capacity throughout

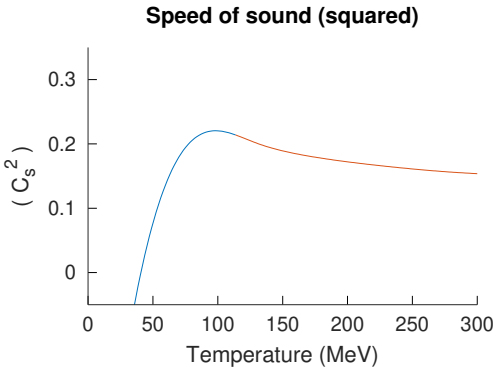


FIG. 7: Asymptotic behavior of Speed of sound(sqaured)

V. ACKNOWLEDGEMENTS

We are grateful to Prof. R. Ramanathan (Rtd.), Department of Physics and Astrophysics, University of Delhi, for his vision and unparalleled guidance throughout the course of this study. Further, we would like to thank Delhi Technological University, Dr. M. Jayasimhadri from the Applied Physics Department and Kirori Mal College, University of Delhi for the immense support from these institutions. We have used publicly available open-source software GNU-Ocatve [16] for the generation of the plots depicted in this manuscript.

References

[1] R. Ramanathan, Y. K. Mathur, K. K. Gupta, and A. K. Jha, "Simple statistical model for analysis of quark-gluon plasma droplet (fireball) formation," *Physical Review C*, vol. 70, no. 2, Aug 2004.

[2] R. Ramanathan, A. K. Jha, K. K. Gupta, and S. S. Singh, "Effect of finite chemical potential on qgp-hadron phase transition in a statistical model of fireball formation," *Indian Journal of Physics*, vol. 85, no. 6, p. 885–889, Jun 2011.

[3] O. Soloveva, P. Moreau, and E. Bratkovskaya, "Transport coefficients for the hot quark-gluon plasma at finite chemical potential μ_B," *Phys. Rev. C*, vol. 101, p. 045203, Apr 2020. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevC.101.045203

[4] L. Yan and X.-G. Huang, "Dynamical evolution of magnetic field in the pre-equilibrium quark-gluon plasma," 2021.

[5] V. V. SKOKOV, A. Y. ILLARIONOV, and V. D. TONEEV, "Estimate of the magnetic field strength in heavy-ion collisions," *International Journal of Modern Physics A*, vol. 24, no. 31, pp. 5925–5932, 2009. [Online]. Available: https://doi.org/10.1142/S0217751X09047570

[6] G. Cao, "First-order qcd transition in a primordial magnetic field," *Phys. Rev. D*, vol. 107, p. 014021, Jan 2023. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.107.014021

[7] R. A. Abramchuk, M. A. Andreichikov, Z. V. Khaidukov, and Y. A. Simonov, "Dense quark–gluon plasma in strong magnetic fields," *Eur. Phys. J. C*, vol. 79, no. 12, p. 1040, 2019.

[8] A. Peshier, B. Kämpfer, O. Pavlenko, and G. Soff, "An effective model of the quark-gluon plasma with thermal parton masses," *Physics Letters B*, vol. 337, no. 3, pp. 235–239, 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0370269394909695

[9] S. Koothottil and V. M. Bannur, "Thermodynamic behavior of magnetized quark-gluon plasma within the self-consistent quasiparticle model," *Phys. Rev. C*, vol. 99, p. 035210, Mar 2019. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevC.99.035210

[10] S. Somorendro Singh, D. S. Gosain, Y. kumar, and A. K. Jha, "Effect of the curvature on a statistical model of Quark-Gluon-Plasma fireball in the hadronic medium," *Pramana*, vol. 74, pp. 27–37, 2010.

[11] T. Niida and Y. Miake, "Signatures of qgp at rhic and the lhc," *AAPPS Bulletin*, vol. 31, no. 1, May 2021. [Online]. Available: http://dx.doi.org/10.1007/s43673-021-00014-3

[12] K. Fukazawa, T. Muta, J. Saito, I. Watanabe, M. Yonezawa, and M. Inoue, "Dynamical electroweak symmetry breaking with color-sextet quarks," *Progress of Theoretical Physics*, vol. 85, no. 1, p. 111–130, Jan. 1991. [Online]. Available: http://dx.doi.org/10.1143/PTP.85.111

[13] A. Das, H. Mishra, and R. K. Mohapatra, "Electrical conductivity and hall conductivity of a hot and dense quark gluon plasma in a magnetic field: A quasiparticle approach," *Phys. Rev. D*, vol. 101, p. 034027, Feb 2020. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevD.101.034027

[14] G. Huang, J. Zhao, and P. Zhuang, "Quantum color screening in external magnetic field," *Physical Review D*, vol. 107, no. 11, Jun. 2023. [Online]. Available: http://dx.doi.org/10.1103/PhysRevD.107.114035

[15] F. G. Gardim, G. Giacalone, M. Luzum, and J.-Y. Ollitrault, "Thermodynamics of hot strong-interaction matter from ultrarelativistic nuclear collisions," *Nature Physics*, vol. 16, no. 6, p. 615–619, Mar. 2020. [Online]. Available: http://dx.doi.org/10.1038/s41567-020-0846-4

[16] J. W. Eaton, D. Bateman, S. Hauberg, and R. Wehbring, *GNU Octave version 8.4.0 manual: a high-level interactive language for numerical computations*, 2023. [Online]. Available: https://www.gnu.org/software/octave/doc/v8.4.0/

INTERNATIONAL DEVELOPMENT PLANNING REVIEW
ISSN:1474-6743 | E-ISSN:1478-3401

EFFECT OF AGRICULTURAL TRADE LIBERALIZATION ON FARM INCOME IN INDIA: AN EMPIRICAL ANALYSIS USING ARDL APPROACH

Hariom Prakash Verma[1*], Nand Kumar[2]
Delhi Technological University, Bawana Road, Delhi, India-110042

Abstract
This study examines cointigration between farm income and agricultural export using ARDL approach during trade liberalization era (1991-2021) in India. ARDL is one of the most dynamic unrestricted model in economic literature. We have used ARDL approach to account for the non stationary of the data. Several measures and indicators are used to assess farmers' income in the country. Agricultural GDP can provide an approximate proxy for farm income to some extent, but it has limitations. Agricultural GDP represents the total monetary value of all goods and services produced within the agricultural sector. Farmers' income, on the other hand, is the actual earnings that farmers receive from their agricultural activities. This paper considers only Agriculture component of Gross Domestic Product at constant prices as a proxy for farm income. The ARDL Bound test revealed the existence of a long run relationship among selected variables. The coefficient of error correction term indicates that short run shocks could be adjusted towards their long run equilibrium at a rate of 107 percent per annum. The results of the study show that in the long run, agricultural output is positively influenced by exports, fertilizer use, and yield. Imports do not have a significant long-run impact. In the short run lagged differences of imports and fertilizer use have significant affect on farm income. CUSUM test demonstrates stability in the model and diagnostic tests shows that there is no serial correlation as well. While this study has limitations, it provides valuable insights into the relationships between agricultural exports and farm income in India.

Keywords: Liberalisation, Agricultural trade, Farm income, Bounds test, CUSUM, ARDL

1. Introduction

Several countries have carried out economic reform and followed trade liberalisation over last several decades. The impact of trade liberalisation on farmers' income is a significant and complex topic. Trade liberalisation policies include tariff reductions, removal of non tariff barriers, or participation in trade agreements. Trade liberalisation encourages countries to specialise in goods and services in which they have a comparative advantage, which can potentially increase the income. Changes in the factor endowments, such as land, labour and capital may affect farmers' income. Trade liberalisation typically leads to increased market access for agricultural products; this can increase demand and competition, influencing prices

and income. Open markets may lead to greater price fluctuations, affecting farmers' income stability.

Farmers' income refers to the earnings or revenue generated by individuals or households engaged in agricultural activities. This income is primarily derived from the sale of crops, livestock, and other agricultural products, as well as related activities such as agribusiness, agro-processing, and value addition. Measuring farmers' income in India is a difficult task due to the diversity of agricultural practices, landholding sizes, crops grown, and other factors. Several measures and indicators are used to assess farmers' income in the country: Gross income, Net Income, crop income etc. Agricultural GDP can provide an approximate proxy for farmers' income to some extent, but it has limitations. Agricultural GDP is the value of all goods and services produced within the agricultural sector. Farmers' income, on the other hand, is the actual earnings that farmers receive from their agricultural activities.

Agricultural Gross Domestic Product (GDP) stands as a pivotal indicator of the economic health of a nation's agricultural sector, serving as a proxy for the income generated by farmers. In our study we have consider only Agriculture component of GDP, the determination of agricultural GDP is a complex interplay of various factors, each exerting its influence on the overall productivity and economic output of the agricultural sector. This paper examines the factors affecting agricultural GDP and also looks into the cointegration between variables in the long run, employing Autoregressive Distributed Lag (ARDL) approach. The independent variables in this study include Agricultural Export, Import, Yield, and Fertilizer consumption, each contributing uniquely to the agricultural production and income.

Agriculture plays an important role in the process of economic development of countries. The importance of understanding and analyzing the determinants of agricultural GDP lies in its direct correlation with the income of farmers. As the world grapples with evolving global dynamics, such as climate change, trade liberalization, and technological advancements, it becomes imperative to dissect the contributing factors that influence the economic outcomes of those engaged in agriculture.

One of the important variables in this study is Agricultural Export. The volume and value of agricultural exports not only reflect the competitiveness of a nation's agriculture sector but also influence the income generated by farmers. Changes in global demand, trade policies, and market access can significantly impact the export potential of agricultural products, thereby affecting the overall agricultural GDP. Another key variable in our study is Yield which explains the quantity of agricultural output per unit of land. The cost of inputs on seeds, fertilizers, machinery and labor constitutes a significant component of the production process. Fluctuations in input costs can have a direct impact on the on farmers' income.

This study examines the cointigration between the farm income and agricultural export using ARDL approach during trade liberalization era (1991-2021) in India, and also investigates the long run cointegration among the variables. The section two of the study discusses some of the important studies carried out on the topic; section three discusses data sources and methodology which have been used in this paper/research. Results of the study have been

reported and discussed in section four and finally conclusion of the study and its limitations have been presented in section five.

2. Literature Review

International trade has become one of the most important drivers of economic growth of a nation. The overall effect of free trade or regional cooperation may be positive, negative or indifferent, depending on the volume of trade creation and trade diversion Clausing (2001).

Anwor et al (2013) finds in their study that trade liberalization has an impact on the agriculture sector's and its export sub-sector's performance in Nigeria. Yusida et al. (2021) findings indicate that an increase in food crop imports alone could lead to a decrease in farmer income. However, if this rise in food crop imports is accompanied by increased exports and production in non-agricultural products, it could result in an increase in farmer income. Anderson et al. (2006) found that transitioning to free merchandise trade would lead to several positive outcomes in developing countries, like increase in agricultural output and exports.

Khan et al. (2021) discovered that trade liberalization in Pakistan doesn't result in reducing income inequality in the short term. They observed that trade agreements favoring agriculture, while improving income equality, is due to decrease in urban and non-farm household income [21]. The key finding of Dhar (2023) is that India's agriculture sector faces challenges due to WTO agreements, including constraints on domestic support measures like minimum support prices and restrictions on export subsidies, impacting the implementation of national food security initiatives. Faridi (2012) conducted research on the impact of agricultural exports on Pakistan's economy, analyzing data from 1972 to 2008. Contrary to expectations, the study found that increasing agricultural exports hindered economic growth, while non-agricultural exports were found to stimulate economic development. As a result, the study suggested promoting non-agricultural and textile products to foster economic growth in Pakistan.

The agriculture sector's performance is contingent on multiple things. Several researches have been carried out on the factors influencing the agriculture sector. Some of the references include Mehdi (2011) , Udah and Nwachukwu (2014) , Ahmad and Heng (2012) , Raza and Siddiqui (2014) , Kannan (2011) , Tripathi and Prasad (2009) , Solanki (2017) , Kamat et al. (2007) , Aynalem et al. (2019) . Existing research have not reached a clear consensus on the extent and impact of different factors on agricultural production. This section reviews and presents literature from different countries worldwide followed by Indian studies.

Mehdi (2011) analyzes the factors influencing agricultural growth in Iran from 1970 to 2007 using annual time series data and the ARDL model. The study discovered that human capital and investments in the agriculture sector had a favorable and considerable impact on agricultural value addition. Shita et al. (2019) found that in Ethiopia, agricultural technology adoption, arable land area, and real GDP had a positive effect on agricultural productivity (cereal output per hectare). The study utilized the ARDL technique to analyze data from 1990 to 2016. Udah and Nwachukwu (2014) examined factors influencing agricultural growth in Nigeria using time series data from 1960 to 2010. Using OLS approach, researchers discovered that factor productivity and infrastructural development, had a major impact on the expansion

of the farming industry. A study conducted by Ahmad and Heng (2012) utilized the ARDL model to analyze the factors influencing the increase of Agricultural Productivity in Pakistan. The primary influencing elements in agricultural productivity development were identified as fertilizer, human capital, and agricultural credit.

Raza and Siddiqui (2014) used the Johansen Co-integration Approach to examine the factors influencing agricultural output in Pakistan based on data from 1972 to 2012. The findings indicated that the quantity of tractors, enhanced seeds, tube-wells, water supply, and agricultural workforce have a significant and favorable impact on agricultural production.

Kannan (2011) analyzed the factors influencing agricultural growth in India using annual data from 1967/68 to 2007/08 and the Ordinary Least Squares (OLS) method. The projected findings indicate that irrigation, capital formation, rainfall, and fertilizer usage have a considerable favorable impact on agricultural growth.

Tripathi and Prasad (2009) used the OLS regression model with annual data from 1950-51 to 2005/06 to explore factors influencing agricultural productivity in India. Land, labor, and capital were identified as the primary sources of agricultural output.

Solanki (2017) examined factors influencing farm GDP in India from 1990/91 to 2011/12 using a VAR model. The study indicated that public investment, terms of trade, and minimum support price had a positive impact on agriculture GDP, whereas fertilizer subsidy has a negative effect.

Kamat et al. (2007) utilized the Cobb Douglas Production Function with the OLS method to study the factors influencing agricultural GDP in India from 1970/71 to 2002/03. The institutional credit and support price of wheat have a beneficial impact on agricultural productivity, whereas the net irrigated area has a negative effect. A distinct regression model was used specifically for the post-economic reform period. All factors were deemed negligible.

Chand et al. (2015) in his study found that agricultural income earned after adjusting for input costs and wages has been different in the last three decades. Farm income per cultivator increased by 2.74% a year from 1983–84 to 1993–94 and dropped to below 2% in the next decade. From 2004–05 to 2011–12, per cultivator farm income increased by 7.3% a year. Das et al. (2017) discovered a U-shaped relationship between farm size and farm income, suggesting that there is an optimal farm size for maximizing income, and an inverted U-shaped relationship is found between farm income and farm and off farm diversification indicating that there is an optimal level of diversification beyond which income may decline.

Dev (2020) discussed in his paper that in India while the farm laws offer farmers more choices and potential for better prices, uncertainties persist. Addressing imperfections, developing marketing infrastructure and shifting focus from cereal-centric policies to diversified agriculture are crucial for success, aiming to achieve higher agricultural growth and increased farmers' income.

Singh (2020) in his findings suggest that neoliberal reforms and open economic policies have not translated into improved agricultural development or enhanced food security in India, but rather have exacerbated existing challenges and inequalities.

Kumar et al (2021) investigates the major factors affecting the performance of agriculture sector in India by using ARDL approach. There study concludes that certified seeds, fertilizers'consumption, net irrigated area and pesticide have positive effect on agricultural GDP in the long run

3. Data and Methodology

The research work is based on the secondary data obtained from different sources like, Ministry of Agriculture, the Directorate of Economics and statistics, and Reserve Bank of India, Mumbai. For the research work the annual time series data on Agriculture component of GDP at constant prices, export, import, fertilizer consumption and yield have been taken for the time period from 1991 to 2001.

For any meaningful regression model, variable should be stationary because the regression analysis based on non stationary variables may give spurious results. Order of the integration of variables helps to select the appropriate model in the time series. Therefore, we have conducted the stationarity test to establish the order of integration. The study uses Phillips-Perron and KPSS along with Augmented Dickey Fuller (ADF) test to measure unit root. Although ADF test is most widely used unit root test, PP test is considered more robust for the presence of autocorrelation and heteroscedasticity.

If all the variables are stationary at level 1 that is I(I), it implies that they have unit root and are non stationary. In this case, it is generally recommended to use a cointegration analysis or error correction model (ECM) to account for non stationarity of the variables. Cointegration analysis helps identify long term relationship among variables and the ECM allows modeling short term dynamics and adjusting for any deviation from the long term equilibrium.

The most commonly used model for such cases are Engle-Granger two step cointegration test, the Johnsen cointegration test and ARDL Bound test. These test help to determine if there is stable long-term relationship among I(I) variables, and if so an ECM can be constructed to study their short term dynamics. This study uses ARDL model to examine the relationship between farmers income and export, import, fertilizer consumption and yield .The ARDL approach is flexible and widely used method for modeling relationship among nonstationary time series data.

For examining the relationship among the agricultural output and various inputs the following regression model has been specified and estimated.

$$lnAGDP = \alpha_0 + \beta_1 lnEX + \beta_2 lnIM + \beta_3 lnFTZ + \beta_4 lnYLD + \varepsilon_t \quad (1)$$

In the above equation, the dependent variable (AGDP) represents Agriculture component of GDP.EX stands for the agricultural export and IM is agricultural import.FTZ indicates total fertilizer consumption and YLD is yield.

Further, we estimate the equation (2) for ARDL bounds test Proposed by Pesaaran, Shin and Smith (2001). The study assesses the long run relationship between Agricultural GDP, Export, Import, Fertilizer and Yield. As in our study our dependent variable is I(l), we have estimated the equation two for bounds test cointegration.

$$\Delta lnAGDP = \alpha_0 + \sum_{i=1}^{n_1} \alpha_{1i}\, \Delta lnEX + \sum_{i=1}^{n_2} \alpha_{2i}\, \Delta lnIM + \sum_{I=1}^{n_3} \alpha_{3i}\, \Delta lnFTZ + \sum_{i=1}^{n_4} \alpha_{4i}\, \Delta lnYLD + \beta_1 lnEX + \beta_2 lnIM + \beta_3 lnFTZ + \beta_4 lnYLD + \mu_t \quad (2)$$

F statistic obtained from the regressing above equation is compared with critical value obtain from Pesaran et al (2001) if the value of F statistic is greater than the upper Bound we reject the null hypothesis of no cointegration, and the value of F statistic is smaller than the critical value of the lower bound I(0), there is no cointegration. After confirming the existence of cointegration among the variables based on the bounds test we regress ARDL restricted error correction model as given in equation (3) below.

$$\Delta lnAGDP = \alpha_0 + \sum_{i=1}^{n_1} \alpha_{1i}\, \Delta lnEX + \sum_{i=1}^{n_2} \alpha_{2i}\, \Delta lnIM + \sum_{I=1}^{n_3} \alpha_{3i}\, \Delta lnFTZ + \sum_{i=1}^{n_4} \alpha_{4i}\, \Delta lnYLD + \theta ECT_{t-1} + \mu_t \quad (3)$$

The coefficient of error correction term θ shows the speed of adjustment towards the long run equilibrium and remaining coefficient provides short run dynamics.

4. Results and Discussions

4.1.Descriptive Analysis

Figure1. Shows the trend of AGDP, over the past three decades, the Agricultural Gross Domestic Product (GDP) in India has shown both positive and negative trend. In the early

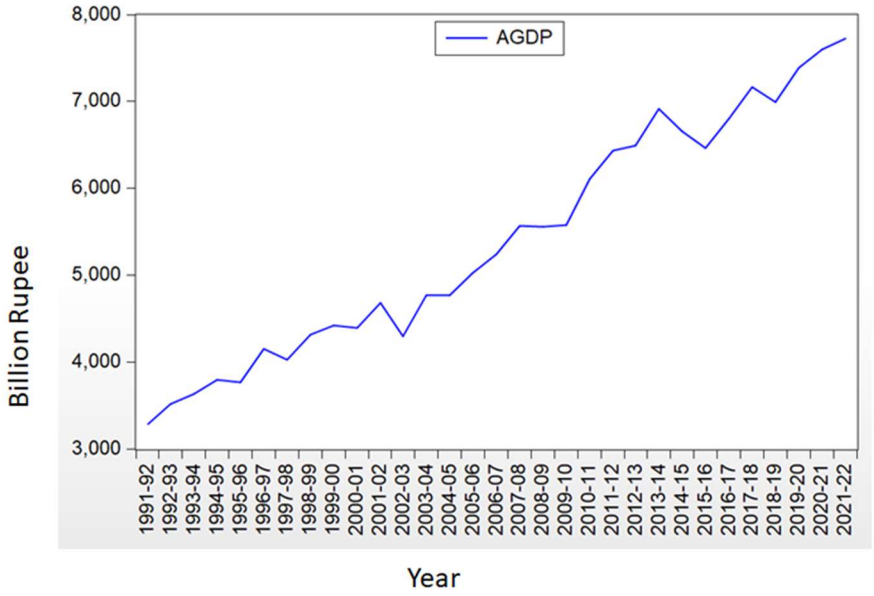


Figure 1. Trend of AGDP

1990s, India faced significant economic reforms aimed at liberalizing various sectors including agriculture. This period witnessed moderate growth in agricultural GDP as the sector started to adapt to new policies and technologies. During 2000 – 2010 the agricultural sector in India saw mixed growth. Factors such as fluctuations in monsoon rains, increased use of technology in farming practices, and government initiatives like the National Agriculture Policy

influenced agricultural GDP growth. Overall, the growth rate was relatively stable but not very high.

The agricultural GDP growth during 2010 – 2020 was somewhat volatile. There were periods of good growth driven by favorable monsoon rains, increased government spending on agricultural infrastructure, and the introduction of schemes like the Pradhan Mantri Kisan Samman Nidhi (PM-KISAN). However, challenges such as agrarian distress, water scarcity, and price fluctuations persisted, impacting overall growth.

The agricultural sector faced significant challenges and transformations during the early years of 2020s, including the COVID-19 pandemic. Despite disruptions caused by the pandemic, the agricultural sector demonstrated resilience, with the government implementing various measures to support farmers and ensure food security.

India embarked on economic reforms in 1991 that liberalized its economy, including the agricultural sector. This period witnessed a gradual increase in agricultural exports as the country started to engage more actively in international trade. Key agricultural commodities exported during this time included rice, wheat, spices, tea, and cotton.

Figure2. shows the trend of Agricultural Export, after 2010 India's agricultural exports saw significant growth in commodities included rice, buffalo meat, spices, fruits, vegetables, and processed food products. The government introduced schemes like the Agricultural Export Policy 2018 to further boost agricultural exports. The COVID-19 pandemic disrupted global trade patterns and supply chains, leading to fluctuations in export volumes and prices. However, India's agricultural exports showed resilience, with certain commodities witnessing increased demand in international markets. The government implemented measures to support agricultural exporters and mitigate the impact of the pandemic on trade.

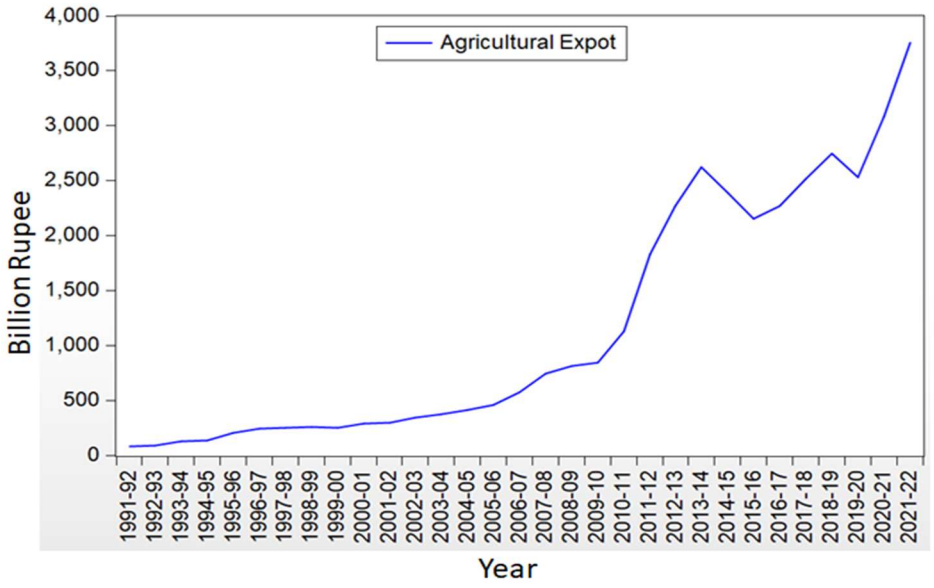


Figure 2. Trend of Agricultural Export

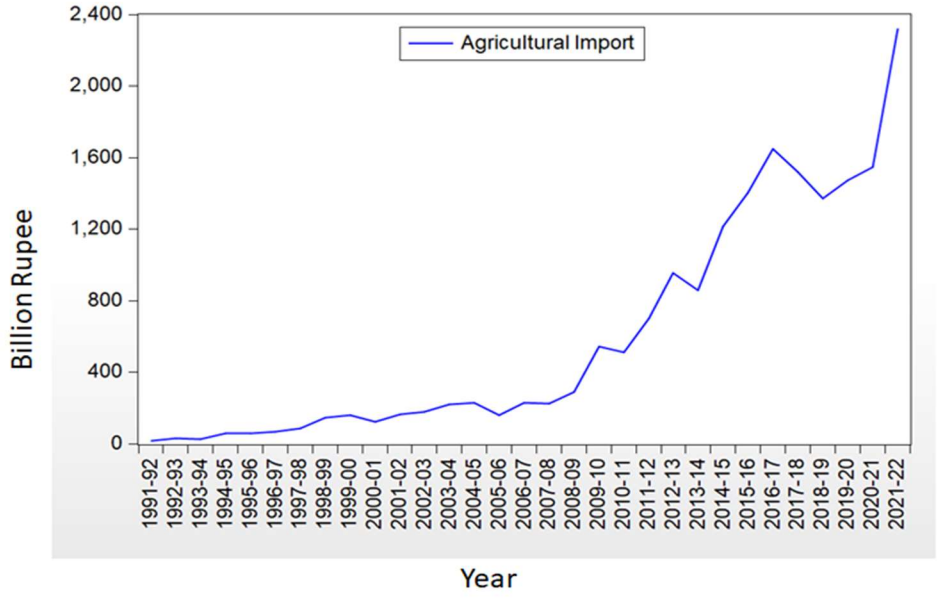


Figure 3. Trend of Agricultural Imports

In the early 1990s, the pattern of agricultural imports underwent significant changes, influenced by shifts in policies, consumer preferences, production capabilities, and global trade dynamics. The general trend over the last 30 years has been one of increasing agricultural imports in India as shown in figure 3.

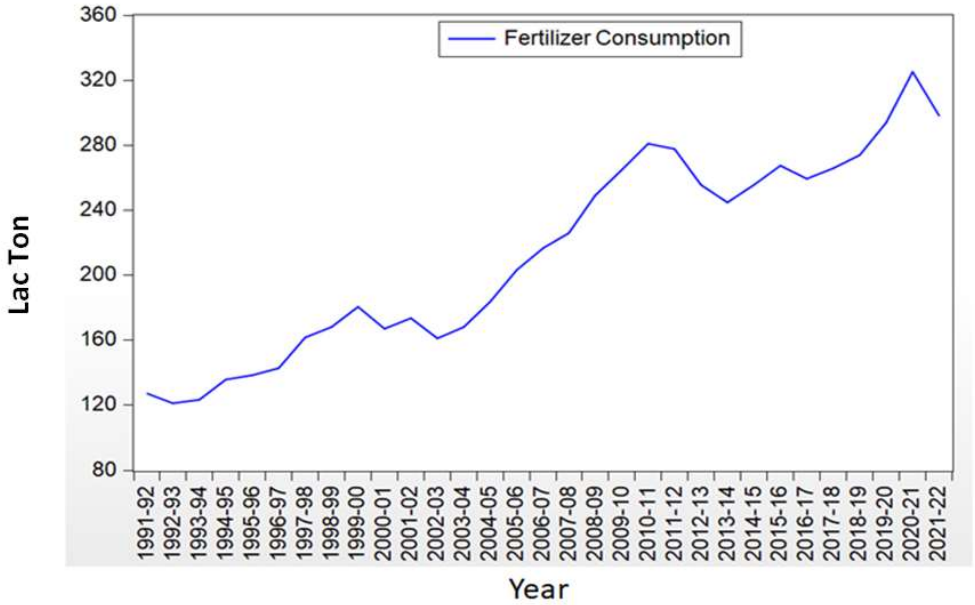


Figure 4. Trend of Fertilizer Consumption

Fertilizer consumption in India has witnessed significant growth due to various factors such as increasing population, expanding agricultural practices, government subsidies, and the adoption of modern farming techniques. Over the past thirty years fertilizer consumption in India has generally trended upwards as shown in figure 4.

Major cereal crops such as rice, wheat, and maize are crucial components of India's food security and agricultural economy. The trend of Yield has shown in figure 5, during 1990s, cereal yields in India experienced moderate growth. The Green Revolution, which began in the 1960s and 1970s, continued to have a significant impact on cereal production during the 1990s. The adoption of high-yielding crop varieties, along with the use of fertilizers, irrigation, and other modern agricultural practices, contributed to yield increases.

India saw a further increase in cereal yields during the 2000s. Government initiatives, such as subsidies for fertilizers, improved access to credit for farmers, and investments in agricultural infrastructure, helped sustain yield growth. Additionally, advancements in

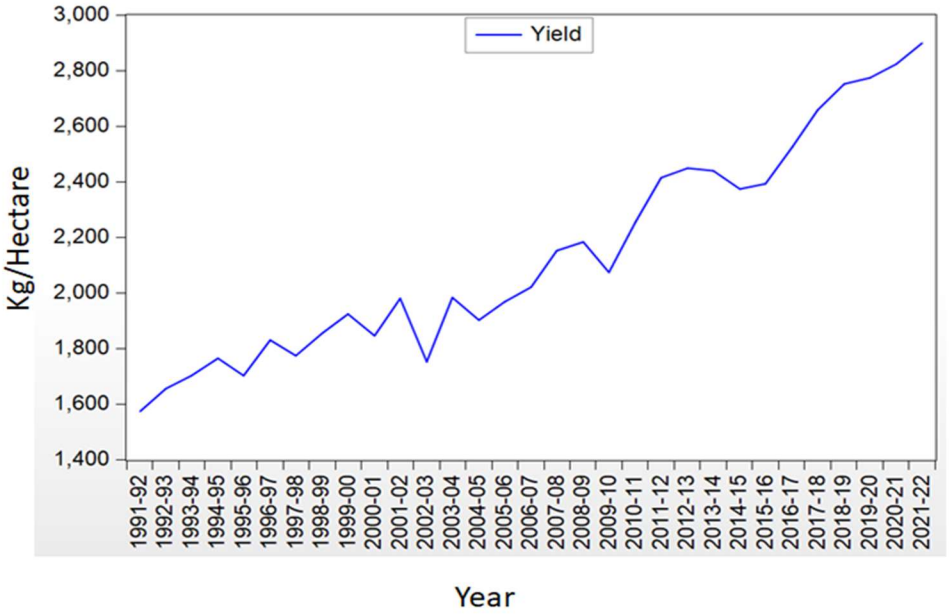


Figure 5. Trend of Yield

agricultural technology and research played a role in boosting productivity. The growth in cereal yields continued during 2010-2020 at a slower pace compared to previous decades. Factors such as climate change, land degradation, and water scarcity posed challenges to agricultural productivity. However, efforts to promote sustainable agricultural practices, including the adoption of precision farming techniques and the use of drought-resistant crop varieties, helped mitigate some of these challenges.

4.2.Econometric Results

4.2.1. Unit Root Test

The Augmented Dicky Fuller Test (ADF) and Phillips Perron (PP) test were used to check the stationarity of the time series data. We have used the ADF and PP tests to check the stationarity of our time series data, examining both levels and first differences. Eviews software has been utilized for this task. The results of the test are reported in the table 1.Based on the results, it is evident that all variables used in the study are stationary at the first difference I(I) and non-stationary at level I(0), according to both the ADF and PP tests. With this information, we have used ARDL to investigate relationships among different variables in the study.

4.2.2. Bounds Test for Cointegration

The ARDL Bound Test, is employed to determine the presence of long-run relationships between two or more variables. This step is necessary to confirm the existence of long-run relationships. The selection of ARDL (2,0,2,2,0) was done using the Akaike Information Criteria (AIC), which is a measure for lag length selection. This selection is done automatically and helps in choosing the most suitable lag lengths based on likelihood considerations. Calculation reveals that the F statistic value is 23.70200 , which exceeds the upper bound critical value at all significance levels. This implies that there is a long-run relationship between the variables in the model.

Table 1: Different Unit Root Test Results

Variable	ADF		PP	
	level	First Difference	level	First Difference
AGDP	-0.91415	-9.206681	-1.2600	-9.432532
	(0.7691)	(0.0000)***	(0.6346)	(0.0000)***
EX	-1.0308	-3.9482	-1.114958	-3.833078
	(0.7285)	(0.0052)***	(0.6966)	(0.0069)***
IM	-0.701163	-1.942669	-1.837684	-7.960116
	(0.8281)	(0.0515)**	(0.3559)	(0.0000)***
FTZ	-0.995100	-4.493620	-0.997600	-4.436147
	(0.7420)	(0.0013)***	(0.7412)	(0.0015)***
YLD	1.029053	-9.704511	0.047622	-19.84993
	(0.9956)	(0.00000)***	(0.9558)	(0.0001)***

Notes: P value is denoted in parenthesis

** and *** denotes rejection of null at 5 percent and 1 percent significance level respectively

Table 2: ARDL Bound Test Cointegration Results

Test Statistic	Value	Signif.	Lower Bound	Upper Bound
F-statistic	23.70200	10%	2.525	3.09
k	4	5%	3.058	3.49

1%	4.28	5.84

4.2.3. Long Run Coefficient

After confirming the cointegration among the variables through the Bound Test, we will now analyze the long-run determinants of Agricultural GDP. The results indicate that in the long run, agricultural output is positively influenced by exports, fertilizer use, and yield. Imports do not have a significant long-run impact.

Table 3: Long Run Relationship among Variables (ARDL Bound)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8.642191	1.716785	5.033939	0.0001
LN_AGDP(-1)	-1.076991	0.154996	-6.948508	0.0000
LN_EXPORT	0.095001	0.028731	3.306525	0.0039
LN_IMPORT(-1)	0.010687	0.018094	0.590624	0.5621
LN_FERTILIZER(-1)	0.267970	0.058688	4.566006	0.0002
LN_YIELD	0.387987	0.086453	4.487863	0.0003

4.2.4. Short Run Coefficient

The short-run dynamic model, represented by equation 3, has been estimated, and its results are reported in the in the table 4. The ECM results indicate significant short-term relationships between agricultural output and the lagged differences of imports and fertilizer use. The error correction term is highly significant, suggesting that any deviation from the long-term equilibrium is corrected quickly in the next period.

4.2.5. Diagnostic Tests for ARDL Model

In conclusion, diagnostic examinations for the ARDL model, including autocorrelation, normality, and stability tests, were conducted, and the results suggested that the estimated model is appropriate. Table 5 illustrates that there is no indication of serial autocorrelation or heteroscedasticity. The Jarque-Bera test verified that the error terms conform to a normal distribution, and Ramsey's RESET test supported the correct specification of the functional form.

Table 4: Short Run Relationship among Variables (ARDL-ECM)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(LN_AGRI(-1))	-0.128120	0.067632	-1.894365	0.0744
D(LN_IMPORT)	-0.020021	0.011921	-1.679448	0.1103
D(LN_IMPORT(-1))	-0.032523	0.011881	-2.737291	0.0135
D(LN_FERTILIZER)	0.225583	0.052246	4.317742	0.0004
D(LN_FERTILIZER(-1))	-0.137095	0.056989	-2.405653	0.0271

CointEq(-1)*	-1.076991	0.079894	-13.48018	0.0000

Table 5: Residual Robustness

Diagnostic tests	Null Hypothesis	Statistics	Probability	Decision
SerialCorrelation (Breusch-Godfrey Test)	No serial correlation	0.6420	0.43	Accept Null Hypothesis
Heteroskedasticity (White Test)	Homoscedastic	0.9922	0.52	Accept Null Hypothesis
Normalitiy (Jarque-Bera Test)	Normally distributed	0.1640	0.92	Accept Null Hypothesis
Model Specification (Ramsey Reset Test)	Functional form is correctly specified	1.4941	0.17	Accept Null Hypothesis

The stability of the model is validate with the help of CUSUM test, as shown in figures 6 and 7.

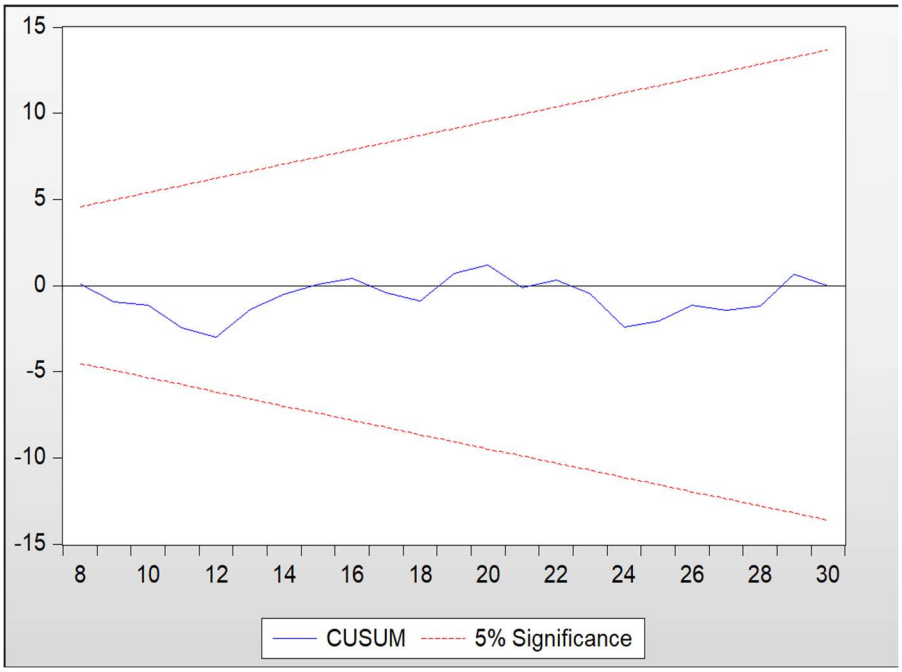


Figure 6. CUSUM Test for Stability

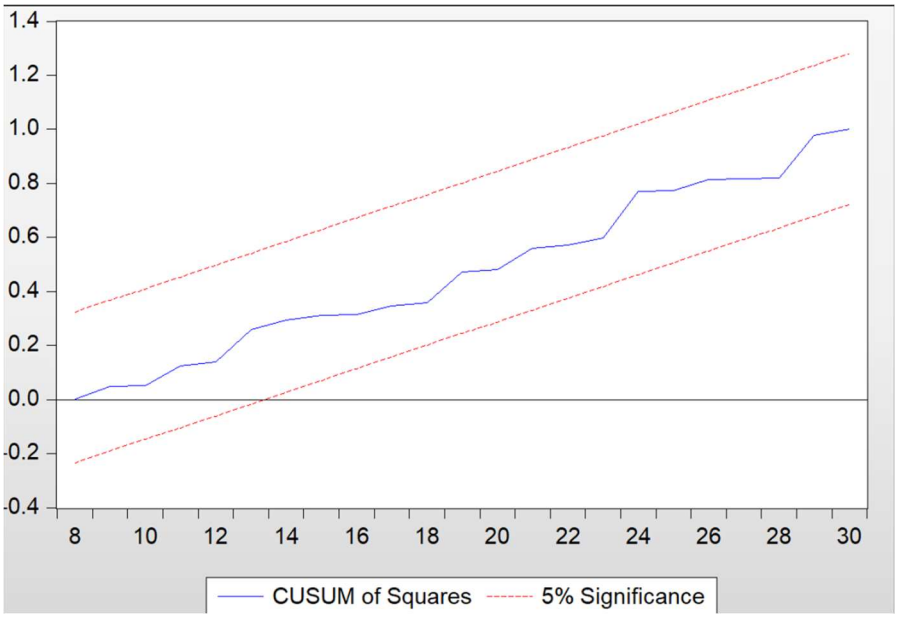


Figure 7. CUSUM Square for Stability

5. Conclusion

In order to understand the impact of trade liberalization on farm's income, this study examined the relationship between farm income and agricultural exports in India during the period of trade liberalization (1991-2021) using the ARDL approach. The ARDL Bound test confirmed the presence of a long-run relationship between farm income and agricultural exports in India. This suggests that changes in agricultural exports have a sustained impact on farm income over time. In the short run, the study identified that lagged differences of Imports and fertilizer use significantly affect farm income.

The results indicate that agricultural exports have a positive and significant effect on farm income in the long run. This underscores the importance of international trade in contributing to the economic well-being of farmers in India. Diagnostic tests, including the CUSUM test, demonstrated the stability of the model. This indicates that the relationships identified in the study are robust and reliable for analyzing the dynamics between agricultural exports and farm income. The study utilized the agriculture component of Gross Domestic Product (GDP) at constant prices as a proxy for farm income. While this approach has limitations, it provides valuable insights into the trends and relationships between agricultural output and farmers' earnings. While this approach has limitations, it provides valuable insights into the relationships between agricultural exports and farm income

DATA AVAILABILITY STATEMENT

The data set used in the manuscript has been obtained from the websites listed below.

1. https://desagri.gov.in/document-report-category/agriculture-statistics-at-a-glance/
2. https://www.worldbank.org/

3. https://www.rbi.org.in/

References

Ahmad, K., & Heng, A. C. T. (2012), Determinants of Agriculture Productivity Growth in Pakistan Determinants of Agriculture Productivity Growth in Pakistan, International Research Journal of Finance and Economics, Vol. 95, pp. 163–172.

Anderson, K., Martin, W., & Van der Mensbrugghe, D. (2006). Distortions to world trade:impacts on agricultural markets and farm incomes. Applied Economic Perspectives and Policy, 28(2), 168-194.

Anowor, O. F., Ukweni, N., & Ikeme, M. (2013). The impact of trade liberalisation on Nigeria agricultural sector. Journal of Economics and Sustainable Development, 4(8), 14-24.

Beckman, J., & Schimmelpfennig, D. (2015). Determinants of farm income. Agricultural Finance Review, 75(3), 385-402.

Chand, R., Saxena, R., & Rana, S. (2015). Estimates and analysis of farm income in India, 1983-84 to 2011–12. Economic and Political Weekly, 139-145.

Clausing, K. A. (2001). Trade creation and trade diversion in the Canada–United States free trade agreement. Canadian Journal of Economics/Revue canadienne d'économique, 34(3), 677-696.

Das, V. K., & Kumar, A. G. (2017). Drivers of farmers' income: The role of farm size and diversification. Indian Agricultural Markets and Policy, 1(1).

Dev, M. (2021). Agricultural reforms in India. Indian Public Policy Review, 2(1 (Jan-Feb), 16- 28.

Dhar, B. (2023). WTO agreement on agriculture: Worsening India's agrarian crisis. The Indian Economic Journal, 71(1), 152-161.

Faridi, M. Z. (2012). Contribution of agricultural exports to economic growth in Pakistan. Pakistan Journal of Commerce and Social Sciences (PJCSS), 6(1), 133-146.

Kannan, E. (2011), Total Factor Productivity Growth and Its Determinants in Karnataka Agriculture, Working paper 265, The Institute for Social and Economic Change, Bangalore.

Kamat, N., Tupe, S., & Kamat, M. (2007), Indian Agriculture in the New Economic Regime, 1971-2003: Empirics based on the Cobb Douglas Production Function, MPRA, Paper No. 6150.

Khan, M. A., Walmsley, T., & Mukhopadhyay, K. (2021). Trade liberalization and Income inequality: The case for Pakistan. Journal of Asian Economics, 74, 101310.

Kumar et al (2021) "Investigating post Economic Reforms Determinants of Agriculture Growth in India by using ARDL approach" IASSI Quaterly, vol. 40, N0. 3,pp. 409-422

Mehdi, S. (2011), Determinants of Agricultural Sector in Developing Countries: The Case of Iran, Annals of Biological Research, Vol. 6, No. 2, pp. 42-55.
Pesaran, M. H., Shin, Y., & Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. Journal of applied econometrics, 16(3), 289-326.

Raza, J., & Siddiqui, W. (2014), Determinants of Agricultural Output in Pakistan: A Johansen Cointegration Approach, Academic Research International, Vol. 5, No. 4, pp. 30-46.

Shita, A., Kumar, N., & Singh, S. (2019), The Impact of Technology Adoption on Agricultural Productivity in Ethiopia: ARDL Approach, Indian Journal of Economics & Business, Vol. 9, No. 1, pp. 255-262.

Singh, P. (2020). Indian agriculture under open economic regime: Implication for livelihood and food security. Journal of Asian and African Studies, 55(8), 1176-1193.

Solanki, S. (2017), Determinants of Agriculture GDP Using VAR Model, KAAV International Journal of Economics, Commerce & Business Management, Vol. 4, No. 2, pp. 295-300.

Tripathi, A. & A.R. Prasad, A.R., (2009), Agricultural Development in India since Independence: A Study on Progress, Performance, and Determinants, Journal of Emerging Knowledge on Emerging Markets,Vol. 1, No. 1, pp. 63-92.

Udah, S. C, & Nwachukwu, I.N., (2014), Determinants of Agricultural GDP Growth in Nigeria, International Journal of Agricultural Research and Review, Vol. 3, No. 3, pp. 184-190.

Yusida, E., Santosa, D. B., Huang, W. C., & Putri, R. D. (2021, January). The impact of agriculture trade liberalization on farmers income in Indonesia: applying CGE model. In Conference on International Issues in Business and Economics Research (CIIBER 2019) (pp. 79-88).

Educational Administration: Theory and Practice
2024, 30(5), 13135-13144
ISSN: 2148-2403
https://kuey.net/

Research Article

Evaluating the Environmental Impact of a Home Appliances: A Life Cycle Assessment with a New Green Product Design Framework

Mohd Tayyab[1*], Ranganath M Singari[2], Peer M Sathikh[3]

[1*,2]epartment of Design, Delhi Technological University, India. *mdtayyab7809@gmail.com
[3]School of Art, Design & Media, Nanyang Technological University, Singapore.

Citation: Mohd Tayyab, et al (2024), Evaluating the Environmental Impact of a Home Appliances: A Life Cycle Assessment with a New Green Product Design Framework, Educational Administration: Theory and Practice, 30(5), 13135-13144
Doi: 10.53555/kuey.v30i5.5674

ARTICLE INFO

ABSTRACT

The objective of this research is to verify the proposed innovative green product design (GPD) framework by conducting a case study that compares and contrasts the environmental impacts of two juicers manufactured by different companies but possessing identical power ratings and capacities. In accordance with the ISO 14044 standard, the research examines the complete life cycle of the household appliance by utilizing life cycle inventory data acquired from a prior publication (Shaukat et al., 2021). The validity of the proposed GPD framework is established through a comparison of its outcomes with those of a prior study that also utilized the life cycle assessment method (Shaukat et al., 2021). A life-cycle analysis reveals that the environmental impacts of the two juicers are notably dissimilar. The first juicer's energy consumption and environmental impact are lower in comparison to the second juicer. Furthermore, the proposed GPD framework's performance is comparable to the method described in the cited article. This study introduces a novel GPD framework for assessing the environmental sustainability of other products and enhances knowledge about the environmental impact of home appliances.

Keywords: Life cycle assessment, Citrus juicer, Green Product Design(GPD), Framework, Environmental impact.

I.Introduction:

Background and Motivation

Exponential growth in the use of home appliances is a result of rapid urbanisation. In 2020, the market for home appliances was expected to be worth around 283 USD billion. (Priyanka Bisht 2015).
The production, use, and disposal of these appliances, however, are responsible for a number of environmental problems. The demand for environmentally friendly products has increased as people's awareness of the effects of climate change and the depletion of natural resources has grown. Understanding how these appliances affect the environment throughout their entire life cycle is crucial.

A widely used method for calculating the environmental effects of products and processes is life-cycle assessment (LCA). Since its introduction in the 1990s, it has become widely embraced. (Allen and Shonnard 2001). Using life cycle assessment (LCA), potential effects of a product are determined throughout all stages of production, including raw material extraction, manufacturing, transportation, use, and final disposal. LCA helps to prevent the shifting of environmental impacts from one life-cycle stage to another by providing incredibly detailed information about the various environmental impacts produced by each stage of the product life cycle. (Bohm et al. 2010; Cooper 2005). The environmental performance of various products, including refrigerators, coffee makers, hand dryers, etc., has also been compared by a number of researchers using LCA. The manufacturing processes that are employed in the creation of a product are greatly influenced by its design. Consequently, a product's design plays a role in determining how it will affect the environment. Different manufacturers use various design methodologies to create products with the same functionality (Shaukat et al., 2021).

Copyright © 2024 by Author/s and Licensed by Kuey. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Objectives of this research are:

To verify the suggested framework's accuracy by contrasting its findings with earlier LCA research that was published.

To advance knowledge of how household appliances affect the environment and to recommend the new GPD framework as a helpful resource for assessing other items' environmental sustainability.

To assess how two different manufacturers' citrus juicers affect the environment.

To determine which phase of a juicer's life cycle results in the greatest environmental impact.

II. Literature Review

The application of LCA is regulated by the standards of the International Standards Organization (ISO) 14040 series (ISO 14044:2006, 2014). In general, two varieties of LCAs are employed: those with multiple figures and those with single figures, such as the proposed framework. Multiple-figure LCA computes a product's ecological repercussions across a variety of impact categories. When comparing two or more product systems, calculations are performed for each of these multiple impact categories. This can be a voluminous, costly, and perplexing procedure that surpasses the capabilities of the majority of seasoned designers.

The environmental effects of various products, including air conditioners (Grignon-Massé, Rivière, and Adnot 2011), cooking appliances (Favi et al. 2018), domestic boilers (Vignali 2017), domestic solar water heaters (Uctug and Azapagic 2018), and refrigerators (Ma et al. 2012), have all been examined by several researchers using LCA. Subassemblies of these products have also been examined by some researchers. Elduque et al. (2014), for instance, performed a life cycle analysis on the electronic boards of an induction hob. The following is a summary of a few key studies on the LCA of home appliances.

Pina et al., (2015) performed a thorough life cycle analysis (LCA) of five distinct induction hob generations. They discovered that newer induction hob generations had less of an impact on the environment than older ones. This discovery was attributed to the more recent induction hob models' effective use of materials and components.

In a different study, a domestic cooker hood's LCA was carried out by Bevilacqua et al., (2010) The majority of the environmental effects, they discovered, were caused by the manufacturing and use phases. The electricity mix used during the use phase, they added, significantly affected the outcomes. In order to lessen the negative effects on the environment, their study also included a modified design that used LED lamps and a three-phase motor.

LCA was used by Hawthorne and Ameta (2012) to assess the environmental effects of blenders. They came to the conclusion that during the entire life of blenders, the manufacturing phase caused the most environmental harm. Additionally, they came to the conclusion that the overall environmental impacts were primarily determined by the materials used to make these blenders. Using LCA, Favi et al. (2018) compared two types of stoves sold in Italy: induction and gas. They discovered that the induction hob had greater negative effects on the environment than the gas hob.

Similarly, several other studies have compared and analyzed the environmental impacts of various products and/or processes using LCA. An examination of two distinct streetlight technologies was conducted in rural Lebanon by Tannous et al. (2018). Hischier (2015) analyzed numerous display technologies using LCA. Scharnhorst, Hilty, and Jolliet conducted a comparative analysis of multiple mobile phone networks in their 2006 study. Rubin et al. (2014) investigated different techniques for copper recovery from printed circuit boards using LCA. A comparative analysis of the environmental impacts of packaging alternatives composed of polymers and glass was conducted by Dhaliwal et al. (2014). In order to further comprehend the ramifications, the LCA methodology facilitated the incorporation of supplementary environmental effects associated with health services. Twenty-23 (Hernández-de-Anda et al.).

In conclusion, the literature review demonstrates that LCA has been used to assess the environmental effects of numerous goods and procedures. The overall environmental effects of these products depend on the design, materials, and manufacturing techniques used to create them. Reduced environmental impact can also be achieved through efficient energy use during the usage phase.

This paper compares two different methods of life cycle assessment and also compares juicers made by two different manufacturers that have the same functionality and power rating, adding to the body of existing literature. This study will highlight how these juicers' various environmental performances differ and relate those differences to their various designs.

III.Methodology

In this study, ISO14040 was used to perform LCA of citrus juicer. This section discusses goal and scope, life-cycle inventory, and life-cycle effect assessment method which is a proposed novel GPD framework. The study's assumptions and sources of inventory data are also explained. According to the ISO standard, an LCA study consists of four steps: objective and scope definition, life-cycle inventory compilation, life-cycle assessment, and outcome interpretation. (ISO 2006).

A Life Cycle Assessment (LCA) study commences by establishing a well-defined scope, functional unit, and system boundaries. Subsequently, a comprehensive record of the resources and energy consumed at every phase of the product's life cycle is compiled. The inventory is subsequently utilized to conduct the impact assessment and calculate various environmental impacts, such as global warming potential.

In pursuance with the objectives of this research, the subsequent procedures were executed:
a. Information Collection From previous publication for life cycle inventory compilation.
b. Define and implement the Proposed GPD Framework for home appliance life cycle assessment.
c. Validation and Analysis of Data.

Information Collection From previous publication

- The functional unit for this study was extraction of a half-litre of juice per day. The life of each juicer was assumed to be 3 years (given three-year warranty by each manufacturer) (Shaukat et al., 2021).
- First juicer (J1) was produced in Slovenia and the second juicer (J2) was produced in China (Shaukat et al., 2021).
- It is postulated that the two juicers were conveyed from the port of Dammam to the retail establishment via vehicle after coming via sea. Both juicers are presumed to be disposed of via landfilling (Shaukat et al., 2021).
- Simapro software was used for modelling life-cycle inventory.

Figure 1, demonstrates the system boundary for the LCA study of both juicers, and Tables 1 demonstrate the weight of various material used for both the juicers.

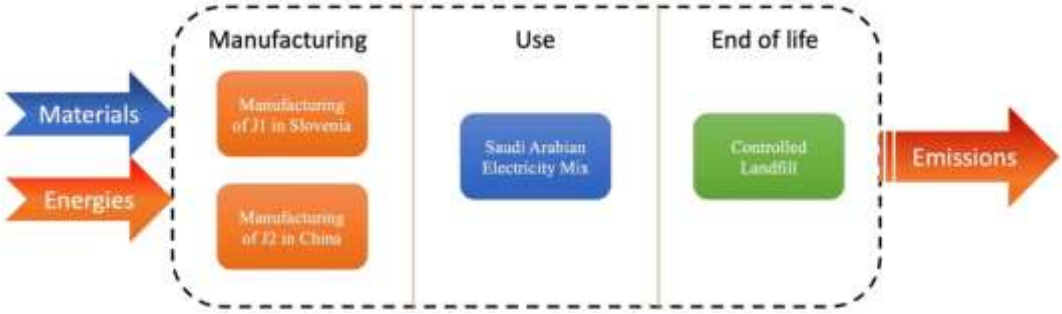


Fig. 1: System boundary for both juicers.

Table1: Life Cycle Inventory for both juicers (Shaukat et al., 2021).

Material	J1	J2
Printed paper (g)	19	12
Cardboard for packaging(g) (corrugated board)	141	117
Polypropylene (g)	336	185
Polyethylene (g) (high density)	29	33
Polystyrene (g)	0	4
Styrene-acrylonitrile (g)	0	186
Steel unalloyed (g)	71	72
Permanent magnet (g)	19	15
Cast Iron (g)	2	2
Copper (g)	27	26
Chromium steel (g)	2	4
Cable (g)	76	114
Transportation, Energy consumption and manufacturing processes		
Transportation by road (kg.km)	18.175	19.275
Transportation by sea (t.km)	9.46	9.885
Electricity usage (MJ)	1.675	2.76
Steel processing (g) (sheet rolling)	69	70
Plastic bag production (g)	0	6
Polystyrene Foam production (g)	0	4
Injection molding of different plastic parts (g)	366	398
Copper wire drawing for motor (g)	26	24

MJ = Mega Joules; g = gram; t.km = tones-kilometre

Proposed GPD Framework

A growing awareness of environmental issues such as material scarcity, high energy consumption, rising atmospheric CO2 levels, ozone depletion, population growth, depletion of natural resources, and so forth contributed to the rise in popularity of "green product design (GPD)" during the late 1980s and early 1990s (Toktas-Palut, P., et al., 2022). The term "sustainable" was initially defined during the Oslo Roundtable in 1994

(Aerais, 2010). Subsequently, a multitude of GPD definitions have been presented, accompanied by illustrations of their application across diverse domains.

Vinodh and Rajanayagam (2010) argue that the main objective of a green production system is to minimize the environmental impact by integrating product and process design with process planning and control. This integration allows for environmental waste flow identification, measurement, evaluation, and management. Green product design provides economic and social benefits to customers, stakeholders, and enterprises, while also serving as a realistic and effective method to counteract environmental degradation (Fuller and Ottman, 2004). The GPD framework also encompasses the concept of eco-design, which aims to harmonize environmental considerations with design objectives that are directed towards economic goals (Karlsson and Luttropp, 2006). Howarth and Hadfield (2006) argue that organizations and designers should prioritize the selection of "environmentally friendly" raw materials. This should be followed by implementing suitable manufacturing and distribution systems, as well as specifying the use and disposal of the final product in a way that minimizes its impact on the environment and society. In the present global context, GPD is extensively advocated as a strategy that can yield substantial advantages. Consequently, numerous scholars have made significant contributions to the advancement and study of unexplored domains within the GPD discipline, with the aim of furthering GPD research.

The proposed Green Product Design (GPD) framework is a design-oriented framework that offers a single-figure score for Life Cycle Assessment (LCA) aspects. This score is significantly more efficient and user-friendly compared to using many categories for LCA. The simplicity of this framework allows for easy implementation using only pencil and paper. Designers who employ this approach develop a deep understanding of the facts, which greatly influences their design process.

This proposed GDP framework (Mohd Tayyab et al., 2024) used similar approach which is used by Philip White (White et al., 2013), which were created with modifications to the TRACI (Tool for Reduction and Assessment of Chemicals and Other Environmental Impacts) impact characterization method that was developed by scientists at the US Environmental Protection Agency (EPA). The factors combine ten environmental impact categories in one single-figure score.

This framework categorizes the product life cycle into three distinct stages: manufacture, usage, and end of life. Figure 2 illustrates this division and provides the carbon emission value for each stage. A lower value indicates a greener product.

The GPD framework can be utilized in several ways to comprehend the environmental efficacy of a design concept, spanning from rapid preliminary assessment to comprehensive examination of the entire product system. Various methodologies can be advantageous, depending on the circumstances and the types of questions the designer aims to address. The primary strategy to utilizing this framework is doing a comprehensive life cycle assessment (LCA) of the entire system. However, it can also be employed for simpler screening purposes and for comparing subassemblies (Mohd Tayyab et al., 2024).

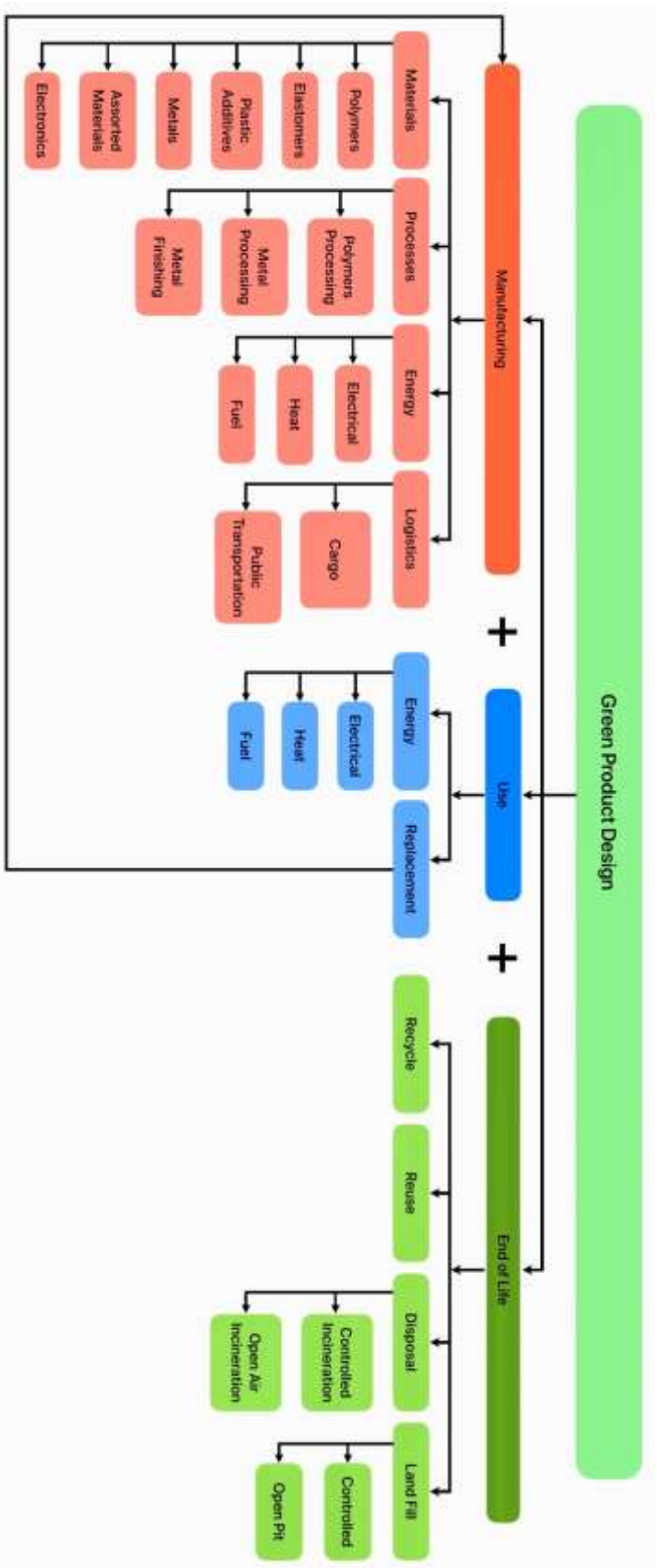


Fig 2: Proposed Framework GPD

The following steps are involved during the life cycle assessment of the product:
Step1: Define lifetime, function unit and system boundaries.
Step2: Make bill-of-materials
Step3: Calculation of estimated impacts

Data Validation and Analysis

The functional unit for this study was extraction of a half-litre of juice per day. The life of each juicer was assumed to be 3 years (given three-year warranty by each manufacturer).
Juicer J1 used an average of 7.5W and J2 used an average of 14W during juice extraction.
Assumed controlled landfill at the end of the life cycle

1. Define lifetime, function unit and system boundaries.

Juicer J1		**Juicer J2**	
Lifetime J1	24 x 365 x 3 = 26,280 hours	Lifetime J2	24 x 365 x 3 = 26,280 hours
Functional Unit	Impacts/life	Functional Unit	Impacts/ life
System boundary	Excludes cleaning during use	System boundary	Excludes cleaning during use

2. Make bill-of-materials

	Juicer J1			**Juicer J2**		
Materials	Printed paper (g)	19	**Materials**	Printed paper (g)	12
	Cardboard for packaging(g) (corrugated board)	141		Cardboard for packaging(g) (corrugated board)	117
	Polypropylene (g)	336		Polypropylene (g)	185
	Polyethylene (g) (high density)	29		Polyethylene (g) (high density)	33
	Polystyrene (g)	0		Polystyrene (g)	4
	Styrene-acrylonitrile (g)	0		Styrene-acrylonitrile (g)	186
	Steel unalloyed (g)	71		Steel unalloyed (g)	72
	Permanent magnet (g)	19		Permanent magnet (g)	15
	Cast Iron (g)	2		Cast Iron (g)	2
	Copper (g)	27		Copper (g)	26
	Chromium steel (g)	2		Chromium steel (g)	4
Manufacturing	Steel processing (g) (sheet rolling)	69	**Manufacturing**	Steel processing (g) (sheet rolling)	70
	Plastic bag production (g)	0		Plastic bag production (g)	6
	Polystyrene Foam production (g)	0		Polystyrene Foam production (g)	4
	Injection molding of different plastic parts (g)	366		Injection molding of different plastic parts (g)	398
	Copper wire drawing for motor (g)	26		Copper wire drawing for motor (g)	24
Use	Electricity usage (MJ)	1.675	**Use**	Electricity usage (MJ)	2.76
Transport	Transportation by road (kg.km)	18.175	**Transport**	Transportation by road (kg.km)	19.275
	Transportation by sea (t.km)	9.46		Transportation by sea (t.km)	9.885
Disposal	landfill		**Disposal**	landfill	

3. Calculations:

Input	**Amount**	**x CO_2 eq** *(kg)*	**= Value** *(kg)*	**Input**	**Amount**	**x CO_2 eq** *(kg)*	**= Value** *(kg)*
Printed paper	0.019(kg)	0.68	0.01292	Printed paper	0.012(kg)	0.68	0.00816
Cardboard for packaging (corrugated board)	0.141(kg)	0.44	0.06204	Cardboard for packaging (corrugated board)	0.117(kg)	0.44	0.05148
Polypropylene	0.336(kg)	1.68	0.56448	Polypropylene	0.185(kg)	1.68	0.3108
Polyethylene (high density)	0.029(kg)	1.50	0.0435	Polyethylene (high density)	0.033(kg)	1.50	0.0495
Polystyrene	0	2.22	0	Polystyrene	0.004(kg)	2.22	0.00888
Styrene-acrylonitrile	0	1.86	0	Styrene-acrylonitrile	0.186(kg)	1.86	0.34596
Steel unalloyed	0.071(kg)	2.13	0.15123	Steel unalloyed	0.072(kg)	2.13	0.15336
Cast Iron	0.002(kg)	0.54	0.00108	Cast Iron	0.002(kg)	0.54	0.00108
Copper	0.027(kg)	1.86	0.05022	Copper	0.026(kg)	1.86	0.04836
Chromium steel	0.002(kg)	2.04	0.00408	Chromium steel	0.004(kg)	2.04	0.00816

Steel processing (sheet rolling)	0.069(kg)	0.16	0.01104
Plastic bag production	0	0.25	0
Polystyrene Foam production	0	1.18	0
Injection molding of different plastic parts	0.366(kg)	0.59	0.21594
Copper wire drawing for motor	0.026(kg)	0.21	0.00546
Transportation by road	18.175 (kg.km)	0.27	4.90725
Transportation by sea	9.46 (t.km)	0.027	0.25542
During Use			
Electricity usage	1.675(MJ)	4.25	7.11875
Controlled landfill			
Printed paper	0.019(kg)	0.43	0.00817
Cardboard for packaging (corrugated board)	0.141(kg)	0.54	0.07614
Polypropylene	0.336(kg)	0.045	0.01512
Polyethylene (high density)	0.029(kg)	0.049	0.001421
Polystyrene	0	0.054	0
Styrene-acrylonitrile	0	0.27	0
Steel unalloyed	0.071(kg)	0.006	0.000426
Cast Iron	0.002(kg)	0.006	0.000012
Copper	0.027(kg)	0.006	0.000162
Chromium steel	0.002(kg)	0.006	0.000012

Steel processing (sheet rolling)	0.070(kg)	0.16	0.0112
Plastic bag production	0.006(kg)	0.25	0.0015
Polystyrene Foam production	0.004(kg)	1.18	0.00472
Injection molding of different plastic parts	0.398(kg)	0.59	0.23482
Copper wire drawing for motor	0.024(kg)	0.21	0.00504
Transportation by road	19.275 (kg.km)	0.27	5.20425
Transportation by sea	9.885 (t.km)	0.027	0.266895
During Use			
Electricity usage	2.76(MJ)	4.25	11.73
Controlled landfill			
Printed paper	0.012(kg)	0.43	0.00516
Cardboard for packaging (corrugated board)	0.117(kg)	0.54	0.06318
Polypropylene	0.185(kg)	0.045	0.008325
Polyethylene (high density)	0.033(kg)	0.049	0.001617
Polystyrene	0.004(kg)	0.054	0.000216
Styrene-acrylonitrile	0.186(kg)	0.27	0.05022
Steel unalloyed	0.072(kg)	0.006	0.000432
Cast Iron	0.002(kg)	0.006	0.000012
Copper	0.026(kg)	0.006	0.000156
Chromium steel	0.004(kg)	0.006	0.000024

Total impact/life Juicer J1: 13.50kg CO_2 eq | **Total impact/life Juicer J2: 18.57kg CO_2 eq**

The process is only accurate to 2 significant figures, so we calculate the impacts per functional unit and round to two significant figures.
*Due to unspecified cable and magnet, the environmental impact of these is not included.

Figure 3(a) and 3(b) shows the environmental impact of both juicer during the three main life cycle stages. It is a graphical representation of proposed GPD framework.

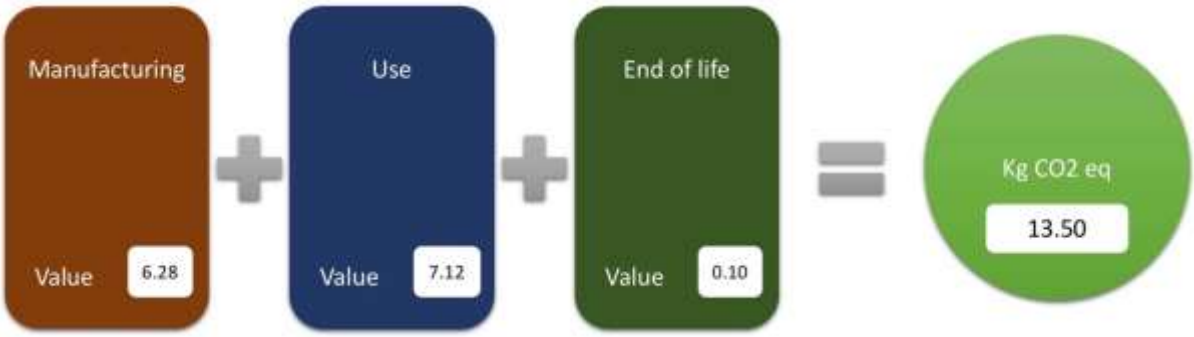


Fig 3(a): Environmental impact of juicer J1 at different stages

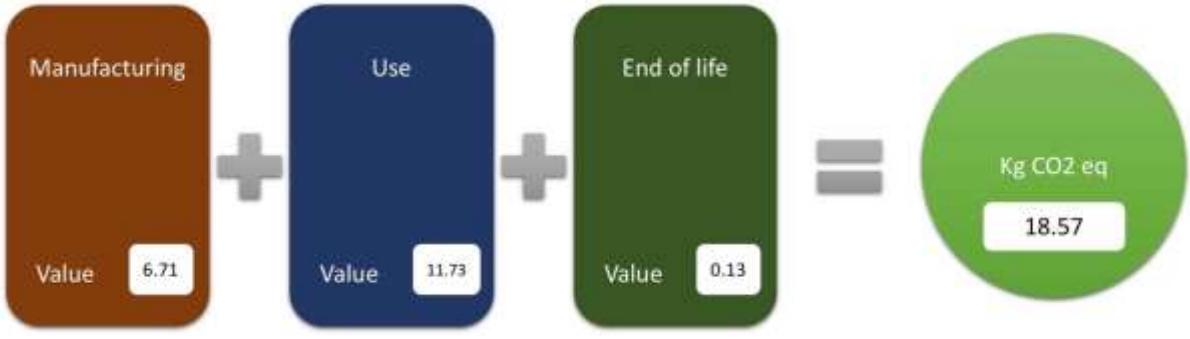


Fig 3(b): Environmental impact of juicer J2 at different stages

IV. Results and discussion

Table 2 presents the characteristics of both juicers, whereas Figure 4 illustrates the respective contributions of different life cycle stages to the environmental impacts of these two juicers. The outcome was determined utilizing the GPD framework and subsequently compared to the outcome of a previously published article that employed the ReCiPe impact assessment technique to ascertain the outcome.

As it is cleared from the result that the juicer J2 which is produced in China having higher environmental impact with compared to juicer J1 which is produced in Slovenia, and it is also cleared from the result that the use phase for both the juicers have majority of environmental impact. On the other hand, the manufacturing phase has second highest environmental impacts during the life cycle of both the juicers, as shown in table 3.

The study also discovered some interesting differences between the two juicers' motors. Compared to the motor used by J2, J1's motor is more effective and uses less power while in use, that's why during use phase juicer J1 impacts lesser than the juicer J2 environmentally.

Table 2: Environmental impact of bicycle at different stages.

Stages	J1 *(Kg CO_2 Eq)*	J2 *(Kg CO_2 Eq)*
During manufacturing	6.28	6.71
During use (for 26,268hr life)	7.12	11.73
At the end of life	0.10	0.13

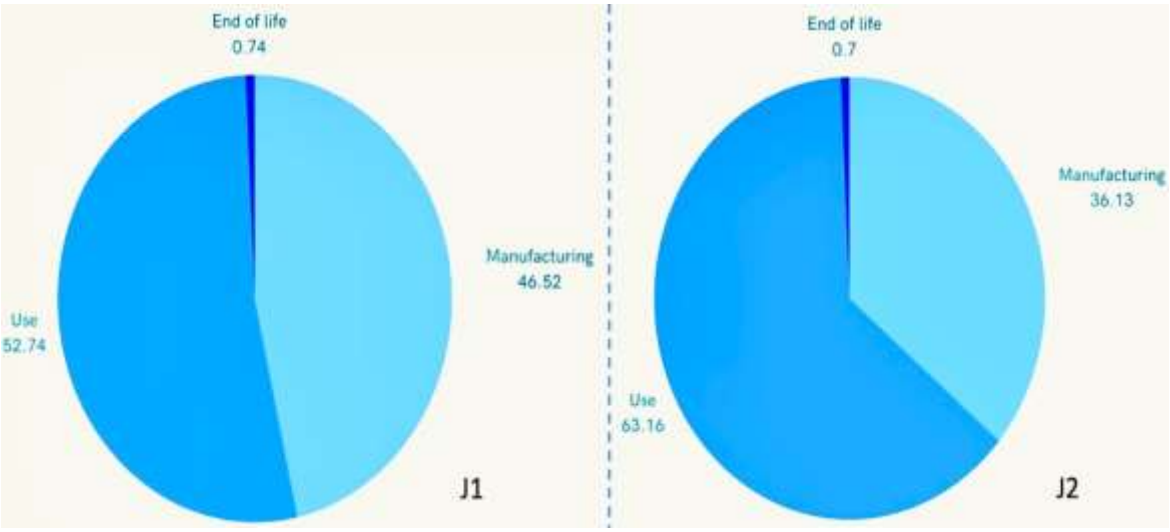


Fig 4: shows the impacts over the entire life cycle of the juicer J1 and Juicer J2.

V. Conclusion

The purpose of this case study was to validate the novel green product design (GPD) framework by assessing the environmental impact of household appliances. The LCI information gathered from a study that was previously published (Shaukat et al., 2021). The research developed a novel GPD framework for life cycle assessment in order to evaluate the ecological consequences of household appliances. Through a comparison of life cycle assessments, it was ascertained that the manufacturing phase had the second-highest environmental impact, trailing only the use phases. The findings revealed that the juicer J2, manufactured in China, exhibited a greater ecological footprint in comparison to the juicer J1, manufactured in Slovenia. This contrast underscores the notion that identically powered and functional products can manifest divergent environmental impacts. The study validated that the results of the revised GPD framework align with those of the previously published framework, indicating that the impacts of transportation and end-of-life phases are inconsequential. The research, which was carried out in Saudi Arabia, encountered obstacles including the scarcity of dependable local datasets, the absence of direct communication with manufacturers, and information regarding supply chain organizations and component transportation from suppliers to final assembly. The results of this study have the potential to contribute to the development of environmentally friendly consumer products by demonstrating that the life-cycle impacts of appliances can be diminished by incorporating energy-efficient motors, optimizing component design, and exercising judicious material selection. This research will help to identify and maintain optimal management and manufacturing practices in the production of household appliances.

References

1. AERAIS (2010), *"Defining green products, white paper, air quality sciences, Atlanta", available at: www.aerias.org/uploads/Defining_Green_Products.pdf (accessed 16 April 2012).*
2. Allen, D. T., and D. R. Shonnard. 2001. "Green Engineering: Environmentally Conscious Design of Chemical Processes and Products." AIChE Journal 47 (9): 1906–1910. doi:10.1002/aic.690470902.
3. Bisht, P. 2015. "Kitchen Appliances Market by User Application (Household, Commercial), Fuel Type (Cooking Gas, Electricity) and Structure (Built-in, Free Stand) - Global Opportunity Analysis and Industry Forecast, 2013-2020 [WWW Document]." Allied Market Research. [Accessed 15 Mar 2018] https://www.alliedmarketresearch.com/kitchen- appliances-market.
4. Bevilacqua, M., Caresana, F., Comodi, G., & Venella, P. (2010, December). Life cycle assessment of a domestic cooker hood. Journal of Cleaner Production, 18(18), 1822–1832. https://doi.org/10.1016/j.jclepro.2010.08.001
5. Bohm, M. R., K. R. Haapala, K. Poppa, R. B. Stone, and I. Y. Tumer. 2010."Integrating Life Cycle Assessment into the Conceptual Phase of Design Using a Design Repository." Journal of Mechanical Design 132 (9). doi:10.1115/1.4002152.
6. Dhaliwal, H., Browne, M., Flanagan, W., Laurin, L., & Hamilton, M. (2014, September 16). A life cycle assessment of packaging options for contrast media delivery: comparing polymer bottle vs. glass bottle. The International Journal of Life Cycle Assessment, 19(12), 1965–1973. https://doi.org/10.1007/s11367-014-0795-1
7. Elduque, D., Javierre, C., Pina, C., Martínez, E., & Jiménez, E. (2014, August). Life cycle assessment of a domestic induction hob: electronic boards. Journal of Cleaner Production, 76, 74–84. https://doi.org/10.1016/j.jclepro.2014.04.009
8. Favi, C., Germani, M., Landi, D., Mengarelli, M., & Rossi, M. (2018, June). Comparative life cycle assessment of cooking appliances in Italian kitchens. Journal of Cleaner Production, 186, 430–449. https://doi.org/10.1016/j.jclepro.2018.03.140
9. Fuller, D.A. and Ottman, J.A. (2004), "Moderating unintended pollution: the role of sustainable product design", Journal of Business Research, Vol. 57 No. 11, pp. 1231-1238.
10. Gámez-García, D. C., Saldaña-Márquez, H., Gómez-Soberón, J. M., Corral-Higuera, R., & Arredondo-Rea, S. P. (2019, January). Life Cycle Assessment of residential streets from the perspective of favoring the human scale and reducing motorized traffic flow. From cradle to handover approach. *Sustainable Cities and Society*, *44*, 332–342. https://doi.org/10.1016/j.scs.2018.10.018.
11. Grignon-Massé, L., Rivière, P., & Adnot, J. (2011, April). Strategies for reducing the environmental impacts of room air conditioners in Europe. Energy Policy, 39(4), 2152–2164. https://doi.org/10.1016/j.enpol.2011.02.004
12. Group M (2018) Bicycle division. Meghna group. http://www.meghnagroup.com.bd/bicycle-division/about-us/. Accessed 2May 2019.
13. Hernández-de-Anda, M. T., Taboada-González, P., Aguilar-Virgen, Q., Velarde-Sánchez, M., & Márquez-Benavides, L. (2023, January). Environmental impacts of a Mexican hemodialysis unit through LCA. Journal of Cleaner Production, 384, 135480. https://doi.org/10.1016/j.jclepro.2022.135480
14. Hischier, R. (2014, October 8). Life cycle assessment study of a field emission display television device. The International Journal of Life Cycle Assessment, 20(1), 61–73. https://doi.org/10.1007/s11367-014-0806-2
15. Howarth, G. and Hadfield, M. (2006), "A sustainable product design model", Materials & Design, Vol. 27 No. 10, pp. 1128-1133.
16. Mohd Tayyab, Ranganath M Singari2,Peer M Sathikh,(2024), E Evaluating the Environmental Impact of a Bicycle: A Life Cycle Assessment with a New Green Product Design Framework., Educational Administration: Theory and Practice, 30(5), 3823-3831
17. Hawthorne, B. D., & Ameta, G. (2011, January 1). LCA Study and Comparison of Two Multispeed Blenders. Volume 2: 31st Computers and Information in Engineering Conference, Parts a and B. https://doi.org/10.1115/detc2011-48612
18. ISO 14044:2006. (2014, August 12). ISO. https://www.iso.org/standard/38498.html.
19. ISO. 2006. "14040: Environmental Management–life Cycle assessment—Principles and Framework." Int. Organ. Stand.
20. Kbah, Z., Erdil, N. O., & Aqlan, F. (2020). Risk Assessment in Oil and Gas Industry Using Simulation and Bow-Tie Analysis. International Journal of Industrial Engineering: Theory, Applications and Practice, 27(1). https://doi.org/10.23055/ijietap.2020.27.1.4042
21. Karlsson, R. and Luttropp, C. (2006), "EcoDesign: what's happening? An overview of the subject area of EcoDesign and of the papers in this special issue", Journal of Cleaner Production, Vol. 14 Nos 15/16, pp. 1291-1298.
22. Kloepffer, W. (2008, February 13). Life cycle sustainability assessment of products. *The International Journal of Life Cycle Assessment*, *13*(2), 89–95. https://doi.org/10.1065/lca2008.02.376.
23. Ma, J., Yin, F., Liu, Z., & Zhou, X. (2012). The Eco-design and Green Manufacturing of a Refrigerator. Procedia Environmental Sciences, 16, 522–529. https://doi.org/10.1016/j.proenv.2012.10.072

24. Pina, C., Elduque, D., Javierre, C., Martínez, E., & Jiménez, E. (2015, April 28). Influence of mechanical design on the evolution of the environmental impact of an induction hob. The International Journal of Life Cycle Assessment, 20(7), 937–946. https://doi.org/10.1007/s11367-015-0890-y
25. Rubin, R. S., Castro, M. A. S. D., Brandão, D., Schalch, V., & Ometto, A. R. (2014, February). Utilization of Life Cycle Assessment methodology to compare two strategies for recovery of copper from printed circuit board scrap. Journal of Cleaner Production, 64, 297–305. https://doi.org/10.1016/j.jclepro.2013.07.051
26. Scharnhorst, W., Hilty, L., & Jolliet, O. (2006, July). Life cycle assessment of second generation (2G) and third generation (3G) mobile phone networks. Environment International, 32(5), 656–675. https://doi.org/10.1016/j.envint.2006.03.001
27. Tannous, S., Manneh, R., Harajli, H., & El Zakhem, H. (2018, June). Comparative cradle-to-grave life cycle assessment of traditional grid-connected and solar stand-alone street light systems: A case study for rural areas in Lebanon. Journal of Cleaner Production, 186, 963–977. https://doi.org/10.1016/j.jclepro.2018.03.155
28. Toktas-Palut, P., Onan, K., Zahid Gürbüz, M., & Gülden-Özdemir, B. (2022). MOSS SOFTWARE: A NEW TOOL FOR MULTI-OBJECTIVE GREEN SUPPLIER SELECTION. International Journal of Industrial Engineering: Theory, Applications and Practice, 29(2). https://doi.org/10.23055/ijietap.2022.29.2.5903
29. Uctug, F. G., & Azapagic, A. (2018, May). Life cycle environmental impacts of domestic solar water heaters in Turkey: The effect of different climatic regions. Science of the Total Environment, 622–623, 1202–1216. https://doi.org/10.1016/j.scitotenv.2017.12.057
30. Vignali, G. (2017, January). Environmental assessment of domestic boilers: A comparison of condensing and traditional technology using life cycle assessment methodology. Journal of Cleaner Production, 142, 2493–2508. https://doi.org/10.1016/j.jclepro.2016.11.025
31. White, St.Pierre, & Belletire. (2013, January 1). *Okala Practitioner Applying Ecological Design* (2nd ed.) [Handbook]. Okala Team. http://www.okala.net/index.html.

Asia-Pacific Financial Markets
https://doi.org/10.1007/s10690-024-09473-8

ORIGINAL RESEARCH

Examining the Dynamics of India's Major Exchange Rates Using Fourier Nonlinear Quantile Unit Root Test

Khyati Kathuria[1] · **Nand Kumar**[2]

Accepted: 15 June 2024
© The Author(s), under exclusive licence to Springer Japan KK, part of Springer Nature 2024

Abstract
The paper empirically investigates the validity of Purchasing Power Parity (PPP) hypothesis for India with its 20 major trading partners using the Fourier non-linear quantile unit root (FNQKS) test. The study uses daily data for the period 1st January 2020–2nd February 2022. FNQKS test supports PPP in 15 out of 20 trading partners of India. The validity of PPP indicates that it is impossible to obtain unbounded gains from arbitrage in these trading partners because of the adjustment process even in the presence of heavy-tailed distributions, mean breaks, and non-linearity. It also indicates that the impact of shocks on the exchange rates is transitory. Therefore, no interventions in these foreign exchange markets need to be made by the relevant authorities. Thus, non-normal distributions, structural breaks, and non-linear mean reversion appear to be key features for adjustment process of exchange rates of these 15 trading partners.

Keywords Exchange rates · Quantile unit root · PPP · Nonlinear · Fourier · India

JEL Classification F14 · F41 · C1 · C22 · C32

1 Introduction

As the foreign exchange market is one of the largest financial markets globally, the pursuit to comprehend and foresee the dynamics of exchange rates and their associated volatilities has become a topic of notable distinction. Specifically in the floating exchange rates regime, practical implications of these dynamics hold for investors,

✉ Nand Kumar
nandkumar@dce.ac.in

Khyati Kathuria
Kathuria.khyati19@gmail.com

[1] Department of Economics, Satyawati College (Evening), University of Delhi, Delhi, India

[2] Department of Humanities, Delhi Technological University, Delhi, India

Published online: 27 June 2024

market analysts, international trade and macroeconomic policymakers (Junior et al., 2019). In the last two decades, most of the empirical work carried out to test the dynamics of the real exchange rates has focused on testing the Purchasing Power Parity (PPP) hypothesis (Pérez-Rodríguez et al., 2009). Purchasing Power Parity (PPP), the concept that real exchange rates revert to a constant mean following a shock is intimately connected with the topic of testing for unit roots in economic time series. If the unit root hypothesis can be rejected in favor of a level stationary alternative, then the real exchange rate is mean reverting and PPP holds (Amara and Papell, 2006)

Policy-makers and researchers are interested in the long-run relationship of purchasing power parity (PPP) for several reasons. First, if the hypothesis holds, it means that the shocks to real exchange rates (RERs) would be temporary. A widely held view is that exchange rate stability promotes foreign trade and investments, and countries tend to benefit from financial markets that are broadening, deepening and internationally integrated. Second, PPP has become a prediction model of exchange rate and a popular criterion for judging currency misalignments. Besides, the validity of the hypothesis provides no room to implement a depreciation policy to deal with current external imbalances. Therefore, it is not surprising to find that much effort has been made in the past two decades on the validity of PPP by using a variety of approaches and sampling periods to resolve the unsettled issue (Soon et al., 2015).

Significant theoretical and empirical advances have taken place in the area of exchange rate economics, however, puzzles relating to determination of exchange rate are still unresolved. A range of vital issues like, the link between exchange rates and economic fundamentals, the purchasing power parity (PPP) puzzle, effects of different exchange rate regimes on output and welfare, relation between exchange rates and global prices etc. have been studied using interesting methodologies. Nonetheless, the continuing economic crisis due to the euro zone crisis, headed by the global financial crisis and lately, the Brexit, indicates that imperative work (both theoretical and empirical) on exchange rates needs to be done (Manzur, 2018).

Several methodologies have been used in the empirical literature for testing long-run validity of PPP, like the univariate unit root and panel unit root tests, multivariate cointegration and panel cointegration tests. The first wave of studies of long-run PPP used univariate unit root tests (e.g. Taylor, 1988). However, low power of these tests over short time spans of data resulted in a failure to find strong evidence of PPP. Therefore, panel unit root tests which exploit both the cross-section and time series variation of the data were used to increase the power of these tests (Lothian & Taylor, 1997; Taylor, 1990). Some authors also used long spans of data to increase the power of these tests (e.g. Lothian & Taylor, 1996). While using long spans of data to raise the power, it also enhances the possibility of structural breaks (Hegwood & Papell, 1998), panel unit root tests have also been criticized by Taylor and Sarno (1998), as the rejection of joint null hypothesis of unit root behaviour in favor of the level stationary alternative does not necessarily provide evidence that all individual currencies exhibit mean reversion. Thus, neither long-span studies nor panel unit root studies are perfect (Doğanlar, 2009).

The empirical evidence on PPP as a long-run equilibrium phenomenon is mixed and the lack of empirical evidence is referred to as the PPP puzzle (Bahmani-Oskooee et al., 2018). To solve the puzzle, an important development in the literature is to allow for structural shifts and nonlinearity in RER because changes in exchange rate regimes, long-lived events and unexpected crashes may lead RER to persistently deviate from its long-run equilibrium and may show an asymmetric adjustment pattern. In order to account for asymmetric adjustment along with structural changes, and non-linearity as well as non-normal distributions, a growing strand of the literature has benefited from quantile unit root inference. While the conventional unit root tests assume a constant speed of adjustment and are able to explain an average behaviour of RER, quantile unit root approaches account for the effects of shocks with various size and sign; identify asymmetric unit root behaviour at quantiles of the distribution; and offer large efficiency gains when the data accommodates potential heavy-tailed (non-normal) distributions. More recent studies have employed quantile unit root tests to account for asymmetric adjustment as well as non-normal distributions (Nazlioglu et al., 2021).

Like all exchange rates of the world, India followed a fixed exchange rate regime until the breakdown of the Bretton Woods system in the early 1970s. However, the exchange rate system remained pegged (first with Pound Sterling and then with a Basket of Currencies) till the dual exchange rate system was introduced in March 1992 for the first time, making the rupee partially convertible. Since March 1993, the dual exchange rate system was merged and the rupee became fully flexible in the current account. The removal of capital account controls aimed to treat inflows and outflows of capital differently: controls on the inflow of capital, especially foreign direct investment, were removed faster than the outflow. This was done to hedge against an enhanced indebtedness of the economy due to rapid capital outflow. Flexibility was combined with Central Bank interventions. Coupled with this, India also undertook a massive liberalization programme since the early 1990s, especially with regard to international trade and capital flows. This has led to a sudden spurt in foreign exchange transactions at a time when the exchange rates were far freer to fluctuate than what they were a decade ago. With the possibility of fluctuations, speculative trading in exchange rates also increased manifold during this period, making their markets complex and exposed to a potentially large set of information (shocks), both from home and abroad.

It is therefore interesting to study how emerging markets like India that have experienced several structural policy changes and internal and external shocks caused exchange rates to have heavy tails, structural breaks and non-linear dynamics. To the best of our knowledge, only a few previous studies simultaneously account for these features of RER in emerging markets with an asymmetric persistence framework, therefore, this study tries to fill this gap in the literature. Thus, given the theoretical and operational significance of the issue and mixed empirical evidence, the present study empirically investigates the validity of the PPP hypothesis for India with its 20 major trading partners using the Fourier non-linear quantile unit root test proposed by Bahmani-Oskooee et al. (2020) which is robust to non-normal distributions, structural breaks, and non-linearity.

2 Literature Review

The literature on this area can be divided into four groups. The first group of studies is those that use univariate unit root tests to examine PPP by testing for stationarity of exchange rates. The second group of studies is those that use panel unit root techniques to test for stationarity of real exchange rates. The third group of studies examines stationarity of real exchange rates using nonlinear unit root techniques. The fourth group of studies applies cointegration tests on the relationship between nominal exchange rates and the domestic and foreign price levels.

Su et al. (2012a, 2012b) tested the validity of long-run PPP for a sample of BRICS countries (i.e. Brazil, Russia, India, China and South Africa) over January 1996 to July 2010 using both linear (ADF) and nonlinear unit root tests with stationary covariates as proposed by Hansen (1995) and Tsong (2011). Empirical results from both linear and nonlinear unit root tests indicated that PPP was true for all the BRICS countries.

Narayan (2006) examined the stationarity of India's real exchange rate with 16 of its major trading partner countries for the period 1960–2000. Application of the conventional ADF unit root test, the Lagrange multiplier (LM) unit root test with one structural break and the LM unit root test with two structural breaks provided evidence that India's exchange rate with 15 out of 16 countries supported PPP.

Validity of PPP was re-examined by Öge Güney et al. (2012) for some emerging market economies and African countries by extending the recent works of Doganlar et al. (2009) and Chang et al. (2010a, 2010b) respectively. Unit root tests that allowed for nonlinearities and structural change in the data-generating process suggested that although linear unit root test provided evidence in favor of PPP only in a few cases, the new nonlinear unit root tests suggested that the PPP proposition held in a majority of the sample countries.

Amara and Papell (20062013) tested for purchasing power parity in post-Bretton Woods real exchange rate data from 20 developed countries using univariate tests and covariate augmented versions of the Augmented Dickey–Fuller (CADF) and feasible point optimal (CPT) unit root tests. Very strong evidence is found of PPP using the CPT test while much less evidence is found of PPP with the CADF and univariate tests.

Pérez-Rodríguez et al. (2009) performed the out-of-sample real exchange rate forecasts using an Artificial Neural Network model, an Autoregressive model and a random walk model. The results confirmed the relevance of nonlinear adjustment in the dynamics of the real exchange rate.

Using monthly data from January 1970 to October 2011, Karlsson et al. (2017) employed panel unit root tests developed by Taylor and Sarno (1998) and Månsson and Sjölander (2014) to empirically investigate if the RER followed a stationary linear and ESTAR process. A strong support for the PPP in Northeast and Southeast Asia, Central and South America, and sub-Saharan Africa was found.

Su et al. (2012a, 2012b) applied Panel SURKSS tests, which are the Kapetanios et al. (2003) tests based on the panel estimation method of SUR, to test the validity of PPP for 15 Common Markets for Eastern and Southern Africa (COMESA) and/or

the Southern African Development Community (SADC) countries. Results revealed that PPP did not hold for the 15 countries under study.

Bahmani-Oskooee et al. (2015) applied the sequential panel selection method combined with a Fourier function to test the PPP in six major oil exporting countries. The results supported PPP in all six countries except Russia.

Panel Seemingly Unrelated Regressions Augmented Dickey–Fuller (SURADF) tests advanced by Breuer et al. (2001) were used by Chang et al. (2010a, 2010b) to test the validity of Purchasing Power Parity (PPP) for 15 countries belonging to the Common Market for Eastern and Southern Africa (COMESA) and the Southern African Development Community (SADC) from December 1994 to July 2008. The empirical results indicated that PPP did not hold for these 15 countries; however, Breuer et al.'s (2001) panel SURADF tests unequivocally indicated that PPP was valid for Botswana, Madagascar and Malawi.

Chang et al. (2012a, 2012b) applied a panel SURADF test with a Fourier function to investigate the properties of long-run purchasing power parity (PPP) in a sample of transition countries between January 1995 and December 2008. Results from the panel SURADF test with a Fourier function indicated that PPP was valid for all the countries, except Russia.

Bahmani-Oskooee and Ranjbar (2016) applied six different univariate unit root tests to real effective exchange rates of 23 OECD countries. Stationarity or Purchasing Power Parity theory (PPP) was supported only in five countries.

Hooi and Smyth (2007) applied univariate and panel Lagrange Multiplier (LM) unit root tests with one and two structural breaks to real exchange rates for 15 Asian countries. The univariate LM unit root tests found evidence of PPP for two-thirds of the sample. The results from the panel LM unit root test supported long-run PPP for all the Asian countries.

Soon et al. (2015) showed that over the period of January 1986–July 2010 PPP hypothesis held true in most of the US dollar real exchange rates using cointegration method that accounted for breaks in the models and obtained a mean half-life estimate of about 10 months for PPP to converge to its long-run equilibrium level.

Doganlar (1999) investigated the validity of PPP hypothesis using cointegration techniques for five developing Asian countries. However the study could not find any supportive evidence for PPP for India, Indonesia, Philippines and Pakistan. However PPP was supported for Turkey when the Johansen's maximum likelihood approach was used.

Doganlar et al. (2009) empirically investigated PPP for 10 emerging market economies namely Brazil, India, Indonesia, South Korea, Mexico, Pakistan, Peru, Philippines, South Africa and Turkey by using cointegration technique. The exchange rate and price series did not appear to be cointegrated for most of the countries. Also, PPP was rejected for Brazil, India, Indonesia, Pakistan, Philippines, South Africa, South Korea and Turkey but accepted for Mexico and Peru.

Chang et al. (2010a, 2010b) used the momentum threshold cointegration tests to investigate whether any asymmetric adjustment was discernible for BRICS from July 1992 to December 2006 and showed that while the Engle–Granger test (which assumes only symmetric adjustment) failed to reveal any cointegrational

relationship for BRIC the threshold cointegration test (with asymmetric adjustment) provided clear evidence of long-run PPP for BRIC except China.

Chang et al. (2012a, 2012b) used Autoregressive Distributed Lag procedure for threshold cointegration by Li and Lee (2010) to test the validity of long-run PPP for a sample of BRICS countries from January 1996 to July 2010. The empirical results indicated that PPP held true for most of these countries except Brazil.

Bozoklu and Kutlu (2012) analyzed the validity of PPP for eight developing countries by applying Johansen's linear () and Breitung's non-linear (2001) cointegration tests to the monthly dataset from 1983:01 to 2010:06. Results suggested mixed evidence for long-run validity of PPP. For linear cases, the results of Johansen's test were more reliable when compared to Breitung's test. However, for Korea, Malaysia and Thailand results obtained from Breitung's rank test suggested that once the sources of non-linearities were taken into account, stronger evidence on the empirical fulfillment of PPP could be found.

Johansen's cointegration technique was utilized by Kargbo (2009) with annual data from 1955 to 2005 on exchange rates and GDP deflators to investigate the validity of long-run PPP in emerging markets. The study found support for PPP.

Basher and Mohsin (2004) examined the empirical validity of relative PPP doctrine in the context of 10 Asian developing economies from January 1980 to December 1999 by applying a panel cointegration method. The empirical findings of this study did not support the relative version of PPP.

Hasan (2006) empirically examined the Purchasing Power Parity hypothesis using more than a century span of annual data of Australia, Canada and Britain. Within the framework of both linear and non-linear cointegration tests the study found support for the validity of the Purchasing Power Parity hypothesis in the long-run. The results also indicated non-linear mean reversion behaviour in the case of Canada.

Peng et al. (2017) revisited PPP for BRICS countries over 1995M1 to 2015M2 using a quantile unit root test with stationary covariates proposed by Galvao (2009). The empirical findings indicated a support of PPP for all the BRICS countries under study.

Mike and Kızılkaya (2019) tested the long-run validity of PPP by using Fourier quantile unit root and Fourier cointegration analyses for 12 emerging market economies that practiced a flexible exchange rate regime from 2001Q1 to 2016Q4. Results showed that real exchange rate series were stationary for Colombia, India, Philippines, Poland, South Africa, and Tu0rkey. On the other hand, Fourier cointegration test results revealed that PPP was valid for Brazil, Colombia, India, Mexico, South Africa, Thailand, and Turkey.

Bahmani-Oskooee et al. (2017) tried to establish the stationarity of real effective exchange rates of 29 African countries. Conventional linear unit root tests found support for the PPP in eight countries while newly introduced non-linear quantile unit root test found support for the PPP in 15 countries.

Bahmani-Oskooee and Wu (2017) aimed at solving the puzzle by applying the fourier quantile-based unit root test in 34 OECD countries from 1994:01 to 2016:03. While three traditional unit root tests failed to support the PPP

hypothesis, the quantile tests that accounted for sharp shifts and smooth breaks supported the PPP hypothesis in 18 countries.

Oskooee et al. (2017) applied the Quantile unit root test and revisited the Purchasing Power Parity (PPP) in 20 African countries from 1971Q1 to 2012 Q4. While the traditional unit root tests failed to reject unit root hypothesis in most of the countries, Quantile unit root test rejected unit root null hypothesis in at least five countries.

Bahramian and Saliminezhad (2020) employed a Fourier quantile unit root test to examine the PPP hypothesis for the ASEAN-5 economies using the monthly real effective exchange rate data from 1994:01 to 2018:09. While the conventional unit root tests provided conflicting results in support of the PPP hypothesis, Fourier quantile unit root test indicated that the real effective exchange rate series was stationary for Indonesia, the Philippines, Singapore, and Thailand.

Peng et al. (2022) tested the mean reversion properties of the U.S. EPU (economic policy uncertainty) index and its 11 sub-categories using the Fourier nonlinear quantile unit root test over the period 1985M1-2020M12. The results indicated that the U.S. EPU index responded asymmetrically to shocks.

Goswami and Saha (2022) tested the PPP hypothesis for selected 7 cryptocurrencies in 17 countries using the Fourier Nonlinear Quantile Unit Root Test. The results indicated that the PPP was strongly supported when we pooled the data over cryptocurrencies and countries, which is consistent with the PPP testing result for regular currencies.

In this study, while the aim to investigate the evidence of PPP is steady with the extant literature, the methodology is novel. The attention isn't on disentangling proof for PPP for a range of nations rather the emphasis is on one nation, India and its bilateral exchange rate vis-a-vis its trading partners. The novelty of the approach lies in undertaking a detailed investigation of PPP for a particular country. The second innovation pertains to methodology. To this end, the empirical exercise involves a comprehensive treatment of structural changes, non-linearity and non-normal distribution of India's bilateral exchange rates. To achieve this, the study uses Fourier non-linear quantile unit root test proposed by Bahmani-Oskooee et al. (2020a) which is robust to non-normal distributions, structural breaks, and non-linearity.

The plan of the paper is as follows. Section 3 discusses the methodology and sources of data,

Section 4 reports the empirical results, Sect. 5 concludes.

3 Data and Methodology

The study employs daily data from 1st January 2020 to 2nd February 2022 on bilateral exchange rates of India with its 20 major trading partners namely, Australia, Israel, China Mainland, China Hong Kong, USA, UAE, Russia, Japan, UK, Indonesia, Canada, Mexico, Brazil, Malaysia, Turkey, Thailand, Saudi Arabia, Singapore, South Africa and Switzerland. The data is sourced from www.investing.com.

3.1 Methodology

To capture an unknown number of structural breaks of both sharp and gradual nature, Bahmani-Oskooee et al. (2020) encompassed a Fourier approximation into the non-linear quantile unit root test of Li and Park (2018). To save degrees of freedom, Becker et al. (2006) and Enders and Lee (2012) recommend usage of a single frequency to arrive at

$$y_t = \lambda_o + \lambda_1(\tau)\left(\sin\left(\frac{2\pi kt}{T}\right) + \lambda_2(\tau)\cos\left(\frac{2\pi kt}{T}\right) + e_t\right) \quad (1)$$

where y_t is the bilateral exchange rate of India with its trading partners, $\lambda_1(\tau)$ and $\lambda_2(\tau)$ are the coefficients of the fourier terms, e_t is the disturbance term, T is the number of observations and t is a time trend. The reason to include both $\sin\left(\frac{2\pi kt}{T}\right)$ and $\cos\left(\frac{2\pi kt}{T}\right)$ in the model is based on the fact that a Fourier expression is capable of approximating absolutely integrable functions to any desired degree of accuracy. k denotes the number of Fourier frequency. Equation (1) is estimated by OLS for $k\epsilon[0.1,0.2\ldots,5]$ and optimum frequency $k = k^*$ is selected by minimizing sum of squared residuals. Then construct

$$\hat{e}_t = y_t - \hat{\lambda}_o + \hat{\lambda}_1(\tau)\sin\left(\frac{2\pi k^* t}{T}\right) + \hat{\lambda}_2(\tau)\cos\left(\frac{2\pi k^* t}{T}\right) \quad (2)$$

The non-linear quantile autoregression model is described by

$$Q_{\Delta\hat{e}_t}\left(\tau|\widehat{e_{t-1}}\right) = \alpha(\tau)\widehat{e_{t-1}}^3 + \sum_{j=1}^{p}\alpha_j(\tau)\Delta\widehat{e_{t-j}} + \varepsilon_t \quad (3)$$

where $Q_{\Delta\hat{e}_t}\left(\tau|\widehat{e_{t-1}}\right)$ denotes the τ-th quantile of $\hat{e}_t$ conditional on $\widehat{e_{t-1}}$, $\Delta\widehat{e_{t-j}}$ represents the first difference of the estimated error term $\Delta\widehat{e_{t-j}}$ at lag$j$, the cubic term $\widehat{e_{t-1}}^3$ represents a cubic function of the lagged differentiated error term $\widehat{e_{t-1}}$, $\alpha(\tau)$ denotes a parameter or coefficient that varies withτ,$\alpha_j(\tau)$ captures quantile specific effects of the lagged differentiated error terms $\Delta\widehat{e_{t-j}}$, τ represents different quantiles in quantile regression and p is the number of lags. The null hypothesis of unit root $H_0 : \alpha(\tau) = 0$ is tested against the stationary alternative$H_1 : \alpha(\tau) < 0$. The statistic for testing the null hypothesis is defined as

$$t_n(\tau) = \frac{\hat{f}(\widehat{F^{-1}(\tau)})}{\sqrt{\tau(1-\tau)}}\left(\gamma_{-1}{}'P_x\gamma_{-1}\right)^{\frac{1}{2}}\hat{\alpha}(\tau) \quad (4)$$

where $\gamma_{-1} = \widehat{e_{t-1}}^3$, P_x is projection matrix on space orthogonal to $x = (1, \Delta\widehat{e_{t-1}}, \ldots.., \Delta\widehat{e_{t-p}})$, $F^{-1}(\tau)$ denotes the inverse of the cumulative distribution function F evaluated at τ, f represents the estimated regression function and $\hat{f}(\widehat{F^{-1}(\tau)})$ is consistent estimator of $f(F^{-1}(\tau))$. The null hypothesis is also tested over a range of quantiles by Kolmogorov–Smirnov test, given by

$$FNQKS = \sup_{\tau \in [0.10, 0.90]} \left| t_n(\tau) \right| \tag{5}$$

Since $t_n(\tau)$ and FNQKS have non-standard distributions, resampling procedure is used in Bahmani-Oskooee et al. (2020) to obtain empirical distributions of the statistics.

The analysis has been carried out using the econometric software R.

4 Empirical Results

Initially, the preliminary analysis is reported in Table 1. Jarque and Bera (JB) test results indicate non-normal distribution for all the countries except Israel. Ftrig test rejects the null hypothesis of absence of Fourier terms for all the countries in Eq. 1 (i.e. $\lambda_1 = \lambda_2 = 0$), implying significance of smooth breaks. Also, as apparent from Fig. 1 the exchange rate series' are nonlinear. The selected frequency (k^*) is integer in most of the countries and is fractional for eight countries. Integer (fractional) frequency implies temporal (permanent) effects of breaks on the mean of exchange rates (Bahmani-Oskooee et al., 2020).

The preliminary tests reveal the importance of non-linear quantile unit root framework with Fourier approximation. The test has some advantages that make it adequate to control for the stochastic properties of the exchange rate series. First, it is able to control for two types of structural breaks in the data generating process. By using the Fourier expansions to demean the exchange rate series, the unknown numbers, dates, and forms of breaks in the mean are captured. Also, by using the quantile regression to estimate the non-linear ADF regression model, it allows for structural changes in the autoregressive parameters. Second, it allows for a non-linear mean-reverting process within each quantile and heterogeneous speeds of mean reversion across quantiles, which is a specific property that allows for comparing the degree of persistence of positive and negative shocks. Third, the test has good power when the time series variables are distributed non-normally.

However before proceeding with it we carry out a battery of unit root tests. ADF test of Dickey and Fuller (1979) rejects the null of unit root in 12 countries. Fourier ADF (FADF) and Fourier KSS (FKSS) tests (Table 2) capture structural shifts as a smooth process by using Fourier approximation and FKSS further accounts for non-linearity. While FADF test rejects the null hypothesis for 9 countries, FKSS test rejects the unit root in all the 20 countries. Results from Fourier non-linear quantile unit root test (FNQKS) are presented in Table 3. FNQKS test rejects the unit root in 15 out of 20 countries. FNQKS test gives us a general idea about the mean reversion of exchange rate series.

PPP serves as a prediction model for exchange rates and it supports governments with evaluating whether or not the currency is overvalued, which in turn enables a suitable policy response to be specified. However, the possibility of structural breaks and other forms of nonlinearity in exchange rates reflected in the exchange rate dynamics necessitates the need for a more refined analysis to trace the impact of different shock sizes on exchange rate behaviour, something that the Fourier nonlinear

Table 1 Preliminary tests

Countries	JB (*p*-val)	k*	Ftrig (*p*-val)	ADF (*p*-val)		
				Intercept	Trend & Intercept	None
Australia	79.18 (0.000)***	1	662.555 (0.000)***	−1.593 (0.485)	−1.51 (0.825)	0.328 (0.78)
Israel	3.647 (0.161)	1	249.939 (0.000)***	−1.646 (0.458)	−3.761 (0.019)**	1.041 (0.922)
China	29.12 (0.000)***	1	470.709 (0.000)***	−1.531 (0.517)	−3.368 (0.056)*	1.685 (0.978)
Hong Kong	8.891 (0.011)**	1.6	150.733 (0.000)***	−2.966 (0.038)**	−2.923 (0.155)	0.545 (0.833)
USA	11.092 (0.003)***	1.6	173.259 (0.000)***	−2.894 (0.046)**	−2.835 (0.185)	0.553 (0.835)
UAE	10.795 (0.004)***	1.6	176.430 (0.000)***	−2.902 (0.045)**	−2.843 (0.182)	0.55 (0.834)
Russia	133.423 (0.000)***	1	141.997 (0.000)***	−2.938 (0.041)**	−2.886 (0.167)	−1.06 (0.261)
Japan	47.772 (0.000)***	1	408.467 (0.000)***	−2.003 (0.285)	−2.805 (0.195)	−0.131 (0.638)
UK	33.24 (0.000)***	1	905.731 (0.000)***	−1.714 (0.423)	−2.958 (0.145)	0.506 (0.824)
Indonesia	629.941 (0.000)***	2.9	57.224 (0.000)***	−2.63 (0.087)*	−2.896 (0.164)	0.042 (0.696)
Canada	26.942 (0.000)***	1	693.242 (0.000)***	−1.471 (0.547)	−2.879 (0.17)	0.635 (0.853)
Mexico	69.499 (0.000)***	1.1	211.501 (0.000)***	−2.029 (0.274)	−2.537 (0.31)	−0.285 (0.582)
Brazil	390.042 (0.000)***	1.5	100.914 (0.000)***	−3.621 (0.005)***	−3.162 (0.093)*	−1.194 (0.212)
Malaysia	4.773 (0.091)*	1	154.576 (0.000)***	−3.012 (0.034)**	−3.308 (0.065)*	0.262 (0.761)
Turkey	41.459 (0.000)***	1	122.658 (0.000)***	−0.097 (0.947)	−2.044 (0.574)	−2.061 (0.037)**
Thailand	23.975 (0.000)***	1	720.110 (0.000)***	−1.484 (0.541)	−1.966 (0.617)	−0.487 (0.504)
Saudi Arabia	16.612 (0.000)***	1.6	178.295 (0.000)***	−2.987 (0.036)**	−2.928 (0.154)	0.548 (0.834)
Singapore	120.737 (0.000)***	1	200.568 (0.000)***	−2.002 (0.286)	−2.846 (0.181)	0.611 (0.848)
South Africa	25.721 (0.000)***	1.1	942.336 (0.000)***	−1.68 (0.44)	−2.521 (0.317)	−0.323 (0.568)
Switzerland	315.157 (0.000)***	1	110.521 (0.000)***	−2.814 (0.056)*	−2.799 (0.197)	0.771 (0.879)

Number in parenthesis are the *p*-values. ***, **, and * show the significance at 1%, 5%, and 10% levels, respectively

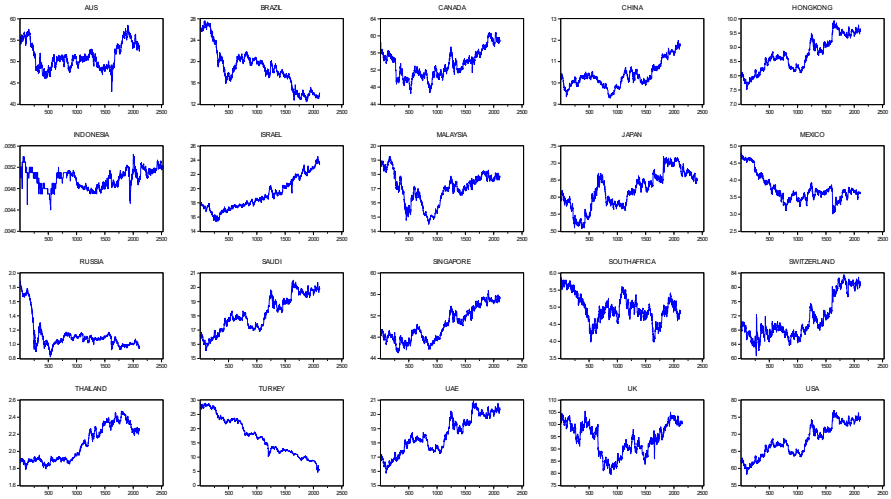


Fig. 1 India's Bilateral Exchange rates with its 20 major trading partners from 1st January 2020 to 2nd February 2022. Author's compilation and data sourced from www.investing.com

Table 2 Fourier ADF and Fourier KPSS test

Countries	FADF				FKPSS			
	Stat	1%	5%	10%	Stat	1%	5%	10%
Australia	−2.814	−4.58	−3.98	−3.67	28.277	0.195	0.127	0.099
Israel	−4.17	−4.8	−4.27	−4	5.038	0.071	0.053	0.046
China	−3.6	−4.8	−4.27	−4	4.546	0.071	0.053	0.046
Hong Kong	−3.684	−4.58	−3.98	−3.67	7.028	0.195	0.127	0.099
USA	−3.883	−4.58	−3.98	−3.67	4.155	0.195	0.127	0.099
UAE	−3.881	−4.58	−3.98	−3.67	4.117	0.195	0.127	0.099
Russia	−3.331	−4.38	−3.75	−3.43	7.983	0.071	0.053	0.046
Japan	−2.904	−4.58	−3.98	−3.67	8.48	0.195	0.127	0.099
UK	−3.592	−4.8	−4.27	−4	4.515	0.071	0.053	0.046
Indonesia	−4.266	−4.8	−4.27	−4	4.706	0.071	0.053	0.046
Canada	−3.978	−4.8	−4.27	−4	5.419	0.071	0.053	0.046
Mexico	−3.88	−4.8	−4.27	−4	4.785	0.071	0.053	0.046
Brazil	−2.608	−4.38	−3.75	−3.43	9.339	0.195	0.127	0.099
Malaysia	−2.845	−4.8	−4.27	−4	3.953	0.071	0.053	0.046
Turkey	−2.96	−4.58	−3.98	−3.67	2.810	0.071	0.053	0.046
Thailand	−3.956	−4.8	−4.27	−4	5.125	0.071	0.053	0.046
Saudi Arabia	−3.908	−4.58	−3.98	−3.67	4.109	0.195	0.127	0.099
Singapore	−4.109	−4.8	−4.27	−4	3.102	0.071	0.053	0.046
South Africa	−3.794	−4.58	−3.98	−3.67	10.136	0.195	0.127	0.099
Switzerland	−4.599	−4.8	−4.27	−4	2.972	0.071	0.053	0.046

Source: Author's computations

Table 3 Fourier nonlinear quantile unit root test

Countries	FNQKS (*p*-val)	*p*-values of $\mathbf{t_n}(\boldsymbol{\tau})$								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Australia	5.865 (0.000)***	0.961	0.769	0.871	0.925	0.844	0.588	0.072	0.001	0.000
Israel	3.331 (0.416)	0.995	0.823	0.869	0.472	0.195	0.338	0.655	0.399	0.106
China	4.027 (0.243)	0.182	0.13	0.066	0.301	0.292	0.264	0.72	0.504	0.493
Hong Kong	5.893 (0.008)***	0.034	0.026	0.003	0.006	0.004	0.079	0.193	0.85	0.947
USA	6.5 (0.003)***	0.057	0.01	0.002	0.001	0.001	0.038	0.12	0.843	0.862
UAE	6.371 (0.003)***	0.019	0.008	0.000	0.001	0.003	0.054	0.117	0.947	0.949
Russia	3.838 (0.068)*	0.921	0.929	0.819	0.763	0.196	0.139	0.01	0.008	0.05
Japan	3.442 (0.106)	0.614	0.358	0.112	0.145	0.073	0.037	0.044	0.211	0.408
UK	4.93 (0.002)***	0.818	0.676	0.258	0.559	0.208	0.003	0.008	0.000	0.005
Indonesia	5.82 (0.000)***	0.806	0.835	0.605	0.189	0.016	0.000	0.000	0.001	0.003
Canada	5.463 (0.094)*	0.926	0.272	0.687	0.15	0.034	0.078	0.282	0.124	0.009
Mexico	8.174 (0.000)***	1.00	0.998	0.911	0.411	0.12	0.001	0.000	0.000	0.000
Brazil	3.391 (0.13)	0.774	0.328	0.127	0.16	0.138	0.038	0.064	0.053	0.024
Malaysia	4.407 (0.005)***	0.488	0.369	0.516	0.304	0.144	0.094	0.055	0.123	0.000
Turkey	5.74 (0.001)***	0.999	1	1	0.998	0.982	0.611	0.321	0.064	0.057
Thailand	5.003 (0.003)***	0.331	0.447	0.335	0.262	0.145	0.039	0.008	0.002	0.032
Saudi Arabia	6.263 (0.009)***	0.147	0.009	0.004	0.000	0.003	0.038	0.132	0.791	0.834
Singapore	3.118 (0.263)	0.612	0.764	0.585	0.464	0.53	0.304	0.506	0.356	0.03
South Africa	4.495 (0.013)**	0.79	0.653	0.224	0.036	0.14	0.039	0.002	0.001	0.103
Switzerland	5.015 (0.002)***	0.97	0.597	0.756	0.454	0.755	0.072	0.000	0.000	0.000

Number in parenthesis are the *p*-values. ***, **, and * show the significance at 1%, 5%, and 10% levels, respectively

quantile unit root test can properly provide. In our results, the validity of PPP in 15 countries indicates that the impact of shocks on the exchange rates is transitory. Therefore, no interventions in these foreign exchange markets need to be made by the relevant authorities (Munir & Kok, 2015). However, contrary to the literature while a shift from traditional or the conventional unit root tests to fourier nonlinear quantile unit root test increased the evidence of satisfaction of PPP hypothesis and thus moved the countries closer to solving the PPP puzzle, in India the same shift led to a reduction in the number of countries satisfying the PPP hypothesis thus leading to the conclusion that the conventional tests (FKSS) overestimated the satisfaction of PPP hypothesis while actually it is satisfied by less number of nations (FNQKS). Therefore in India the evidence from FNQKS test moves us one step backward from solving the PPP puzzle.

However, testing for unit root at specific quantiles for a more detailed analysis of asymmetric persistence can be checked through p-values of t_n statistic at different quantiles reported in Table 3. Exchange rates are stationary at low and /or middle quantiles for 4 countries (Hong Kong, USA, UAE, Saudi Arabia) in which positive shocks have permanent effects on the long run equilibrium. The presence of permanent positive shocks illustrates that the long-run path of exchange rate will be unbounded and thus, these countries can improve its trade balance when its exchange rate increases. Exchange rates are stationary in middle and/or high quantiles in 11 countries (UK, Indonesia, Canada, Mexico, Malaysia, Thailand, Switzerland, Australia, Russia, South Africa and Turkey) implying that negative shocks deviate exchange rates from the long run equilibrium. Thus, due to the long-lasting impacts of the negative and transitory effects of the positive shocks, depreciation has a desirable outcome on trade balance of these nations.

5 Conclusion

This study utilizes a novel approach to test for PPP in exchange rate of India with its trading partners by accounting for non-normal errors, non-linearity and structural shifts. While ADF test support PPP in almost half of the countries, fourier non-linear quantile unit root test supports PPP in 15 out of 20 countries. The validity of PPP indicates that in the long run exchange rates are not influenced by monetary policy, however, official intervention can explain the short-run fluctuations in exchange rates. The satisfaction of PPP can also provide information to international investors and speculators for their portfolio allocations. It also indicates that it is impossible to obtain unbounded gains from arbitrage in these trading partners because of the adjustment process even in the presence of heavy-tailed distributions, mean breaks, and non-linearity. The empirical evidence on the stationarity of exchange rates implies that depreciations could increase international competitiveness and improve trade balance in the short-run, but improvements in trade balance are not unbounded in the long-run.

Thus, the results clearly indicate that non-normal distributions, structural breaks, and non-linear mean reversion appear to be key features for adjustment process of exchange rates of these 15 trading partners. Future studies can re-examine the

international parity conditions (such as the interest rate parity) of India with its trading partners within the context of quantile methods that accommodate these key features. There is also scope for the application of more sophisticated panel data unit root tests to examine the robustness of the findings from the panel unit root tests in this study. The fact that researchers remain divided on the existence of PPP will continue to encourage the development of more sophisticated unit root tests to examine this important topic.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Amara, J., & Papell, D. H. (2006). Testing for purchasing power parity using stationary covariates. *Purchasing power parity and real exchange rates* (pp. 29–39). Routledge.

Bahmani-Oskooee, M., Chang, T., Chen, T. H., & Tzeng, H. W. (2017). Quantile unit root test and the PPP in Africa. *Applied Economics, 49*(19), 1913–1921.

Bahmani-Oskooee, M., Chang, T., Cheng, S. C., & Wu, T. P. (2015). Revisiting purchasing power parity in major oil-exporting countries. *Macroeconomics and Finance in Emerging Market Economies, 8*(1–2), 108–116.

Bahmani-Oskooee, M., Chang, T., Elmi, Z., Gelan, A., & Ranjbar, O. (2018). Non-linear quantile unit root test and PPP: More evidence from Africa. *Applied Economics Letters, 25*(7), 465–471.

Bahmani-Oskooee, M., Chang, T., Niroomand, F., & Ranjbar, O. (2020). Fourier nonlinear quantile unit root test and PPP in Africa. *Bulletin of Economic Research, 72*(4), 451–481.

Bahmani-Oskooee, M., & Ranjbar, O. (2016). Quantile unit root test and PPP: Evidence from 23 OECD countries. *Applied Economics, 48*(31), 2899–2911.

Bahramian, P., & Saliminezhad, A. (2020). Revisiting purchasing power parity in the ASEAN-5 countries: Evidence from the Fourier quantile unit root test. *Applied Economics Letters, 28*(13), 1104–1109.

Basher, S. A., & Mohsin, M. (2004). PPP tests in cointegrated panels: Evidence from Asian developing countries. *Applied Economics Letters, 11*(3), 163–166.

Becker, R., Enders, W., & Lee, J. (2006). A stationarity test in the presence of an unknown number of smooth breaks. *Journal of Time Series Analysis, 27*(3), 381–409.

Bozoklu, S., & Kutlu, S. (2012). Linear and nonlinear cointegration of Purchasing Power Parity: Further evidence from developing countries. *Global Economic Review, 41*(2), 147–162.

Breitung, J. (2001). Rank tests for nonlinear cointegration. *Journal of Business and Economic Statistics, 19*, 331–340.

Breuer, J. B., McNown, R., & Wallace, M. S. (2001). Misleading inferences from panel unit root tests with an illustration from purchasing power parity. *Review of International Economics, 9*, 482–493.

Chang, H. L., Su, C. W., Zhu, M. N., & Liu, P. (2010a). Long-run purchasing power parity and asymmetric adjustment in BRICs. *Applied Economics Letters, 17*(11), 1083–1087.

Chang, T., Lee, C. H., Chou, P. I., & Wang, S. C. (2012b). Purchasing power parity for transition countries: Further evidence from panel SURADF test with a Fourier function. *Eastern European Economics, 50*(4), 42–59.

Chang, T., Lee, C. H., & Hung, K. (2012a). Can the PPP stand on the BRICS? The ADL test for threshold cointegration. *Applied Economics Letters, 19*(12), 1123–1127.

Chang, T., Tang, D., Liu, W., & Lee, C. (2010b). Purchasing power parity for 15 COMESA and SADC countries: Evidence based on panel SURADF tests. *Applied Economics Letters, 17*, 1721–1727.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366a), 427–431.

Doganlar, M. (1999). Testing long-run validity of purchasing power parity for Asian countries. *Applied Economics Letters, 6*(3), 147–151.

Doğanlar, M., Bal, H., & Özmen, M. (2009). Testing long-run validity of purchasing power parity for selected emerging market economies. *Applied Economics Letters, 16*(14), 1443–1448.

Enders, W., & Lee, J. (2012). A unit root test using a Fourier series to approximate smooth breaks. *Oxford Bulletin of Economics and Statistics, 74*(4), 574–599.

Galvao, A. F., Jr. (2009). Unit root quantile autoregression testing using covariates. *Journal of Econometrics, 152*(2), 165–178.

Goswami, G. G., & Saha, T. K. (2022). Fourier nonlinear quantile unit root test of purchasing power parity in cryptocurrencies. *Applied Economics Letters, 31*(4), 312–322.

Hansen, B. E. (1995). Rethinking the univariate approach to unit root testing. *Econometric Theory, 11*, 1148–1171.

Hasan, M. S. (2006). A century of purchasing power parity: evidence from Canada and Australia. *Purchasing power parity and real exchange rates* (pp. 145–156). Routledge.

Hegwood, N. D., & Papell, D. H. (1998). Quasi purchasing power parity. *International Journal of Finance and Economics, 3*, 279–289.

Hooi, L. H., & Smyth, R. (2007). Are Asian real exchange rates mean reverting? Evidence from univariate and panel LM unit root tests with one and two structural breaks. *Applied Economics, 39*(16), 2109–2120.

Johansen, S. (1988). Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control, 12*, 231–254.

Johansen, S. (1994). The role of the constant and linear terms in cointegration analysis of nonstationary time series. *Econometric Reviews, 13*, 205–231.

Junior, P. O., Tweneboah, G., & Adam, A. M. (2019). Interdependence of major exchange rates in Ghana: A wavelet coherence analysis. *Journal of African Business, 20*(3), 407–430.

Kapetanios, G., Shin, Y., & Snell, A. (2003). Testing for a unit root in the nonlinear STAR framework. *Journal of Econometrics, 112*, 359–379.

Kargbo, J. M. (2009). Financial integration and parity reversion in real exchange rates of emerging markets. *Applied Economics Letters, 16*(1), 29–33.

Karlsson, H. K., Månsson, K., & Sjölander, P. (2017). Investigation of the nonlinear behaviour in real exchange rates in developing regions. *Applied Economics Letters, 25*(5), 335–339.

Li, H., & Park, S. Y. (2018). Testing for a unit root in a nonlinear quantile autoregression framework. *Econometric Reviews, 37*(8), 867–892.

Li, J., & Lee, J. (2010). ADL test for threshold cointegration. *Journal of Time Series Analysis, 31*, 241–254.

Lothian, J. R., & Taylor, M. P. (1996). Real exchange rate behaviour: The recent flow from the perspective of the past two centuries. *Journal of Political Economy, 104*, 488–509.

Lothian, J. R., & Taylor, M. P. (1997). Real exchange rate behavior. *Journal of International Money and Finance, 16*, 945–954.

Månsson, K., & Sjölander, P. (2014). Testing for nonlinear panel unit roots under cross-sectional dependency–with an application to the PPP hypothesis. *Economic Modelling, 38*, 121–132.

Manzur, M. (2018). Exchange rate economics is always and everywhere controversial. *Applied Economics, 50*(3), 216–232.

Mike, F., & Kızılkaya, O. (2019). Testing the theory of PPP for emerging market economies that practice flexible exchange rate regimes. *Applied Economics Letters, 26*(17), 1411–1417.

Munir, Q., & Kok, S. C. (2015). Purchasing power parity of ASEAN-5 countries revisited: Heterogeneity, structural breaks and cross-sectional dependence. *Global Economic Review, 44*(1), 116–149.

Narayan, P. K. (2006). Are bilateral real exchange rates stationary? Evidence from Lagrange multiplier unit root tests for India. *Applied Economics, 38*(1), 63–70.

Nazlioglu, S., Altuntas, M., & Kilic, E. (2021). PPP in emerging markets: Evidence from Fourier nonlinear quantile unit root analysis. *Applied Economics Letters, 29*(8), 731–737.

Öge Güney, P., Telatar, E., & Hasanov, M. (2012). Re-examining purchasing power parity for selected emerging markets and African countries. *Applied Economics Letters, 19*(2), 139–214.

Peng, H., Liu, Z., & Chang, T. (2017). Revisiting purchasing power parity in BRICS countries using more powerful quantile unit-root tests with stationary covariates. *Communications in Statistics-Theory and Methods, 46*(20), 10051–10057.

Peng, Y. T., Chang, T., & Ranjbar, O. (2022). Re-Investigating the degree of persistence of US economic policy uncertainty using the Fourier non-linear quantile unit root test. *Applied Economics, 54*(39), 4586–4595.

Pérez-Rodríguez, J. V., Ledesma-Rodríguez, F., & Torra-Porras, S. (2009). Purchasing power parity and nonlinear adjustment. *Applied Economics Letters, 16*(1), 35–38.

Soon, S. V., Baharumshah, A. Z., & Ahn, S. K. (2015). Real exchange rate dynamics in the Asian economies: Can regime shifts explain purchasing power parity puzzles? *Global Economic Review, 44*(2), 219–236.

Su, C. W., Chang, H. L., Chang, T., & Lee, C. H. (2012a). Purchasing power parity for BRICS: Linear and nonlinear unit root tests with stationary covariates. *Applied Economics Letters, 19*(16), 1587–1591.

Su, C. W., Chang, T., & Liu, Y. S. (2012b). Revisiting purchasing power parity for African countries: With nonlinear panel unit-root tests. *Applied Economics, 44*(25), 3263–3273.

Taylor, M. P. (1988). An empirical examination of long-run purchasing power parity using cointegration techniques. *Applied Economics, 20*, 1369–1381.

Taylor, M. P. (1990). On unit roots and real exchange rates: Empirical evidence and Monte Carlo analysis. *Applied Economics, 22*, 1311–1321.

Taylor, M. P., & Sarno, L. (1998). The behavior of real exchange rates during the post- bretton woods period. *Journal of International Economics, 46*, 281–312.

Tsong, C. C. (2011). Testing for a unit root with covariates against nonlinear alternatives. *Economic Modelling, 28*, 1226–1234.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

KSME

Springer

Journal of Mechanical Science and Technology 38 (6) 2024

Original Article
DOI 10.1007/s12206-024-0513-2

Keywords:
· Construction and demolished waste
· Mechanical properties
· Recycled coarse aggregates
· Scanning electron microscopy
· X-ray diffraction

Correspondence to:
Harish Panghal
harish_phd2k18@dtu.ac.in

Citation:
Panghal, H., Kumar, A. (2024). Examining the structural viability of recycled fine aggregates in sustainable concrete. Journal of Mechanical Science and Technology 38 (6) (2024) ?~?. http://doi.org/10.1007/s12206-024-0513-2

Received September 19th, 2023
Revised February 5th, 2024
Accepted March 1st, 2024

† Recommended by Editor
Jun-Hyub Park

© The Korean Society of Mechanical Engineers and Springer-Verlag GmbH Germany, part of Springer Nature 2024

Examining the structural viability of recycled fine aggregates in sustainable concrete

Harish Panghal and Awadhesh Kumar

Department of Civil Engineering, Delhi Technological University, Delhi 1100042, India

Abstract This study investigates the potential of incorporating recycled fine aggregates (RFA) into sustainable concrete. In this research, a conventional compaction technique is utilized to establish the order of compressive strength and, consequently, to assess particle packing density in terms of weight within a specific cylindrical volume and evaluate workability, compressive and flexural strengths, splitting tensile strength, elasticity modulus, and microstructural properties (analyzed through XRD, SEM, and EDAX). The study found that RFA can improve concrete properties, hardened characteristics, and microstructure up to an optimum 25 % RFA replacement threshold (RFA 25). Beyond this value, concrete strength and microstructure deteriorate. RFA 25 exhibits significantly higher compressive (14.75 %), flexural (6.61 %), and splitting tensile (13.14 %) strengths compared with the reference concrete, along with a 5.71 % decrease in the modulus of elasticity. Lower replacement levels promoted pozzolanic reactions, enhancing strength through additional hydration products, whereas higher replacements reduced strength.

1. Introduction

The construction industry, which is crucial for global economic growth, is simultaneously contributing to environmental degradation through the energy-intensive production of concrete, which has a substantial carbon footprint [1]. In addition, the surge in construction and demolition waste (CDW) disposal poses challenges [2], straining landfill space and causing negative ecological impacts [3]. This issue is addressed by incorporating recycled aggregate (RA) from CDW in fresh concrete, constituting 70 %–80 % of its volume, which presents economic and environmental advantages by reusing materials such as concrete, asphalt, bricks and tiles [4-6]. This research focuses on recycled fine aggregates (RFAs) generated through an impact crusher at the IL&FS CDW Recycling Plant in collaboration with the Delhi Metro Rail Corporation [7-9]. However, a comprehensive assessment of its performance and compliance with industry standards are necessary before this substitution is considered in civil engineering projects [10-12]. RA concrete (RAC), which employs RA instead of natural aggregate (NA), has been extensively studied. RAC tends to be less workable due to its porous structure and higher water absorption capacity [13]. Various factors, including the replacement ratio of NA, water-to-cement ratio, parent aggregate type, age, exposure conditions, number of crushing stages, and physical and mechanical characteristics of RA, influence its compressive strength [14-16]. Despite flaws in RAC's microstructure compared with that of NA concrete, adding RAs significantly enhances the strength of hardened concrete, offering a viable alternative to natural sand aggregates with lower permeability [17-19]. Concrete formulations containing RFAs and demolished coarse aggregates have shown effectiveness in replacing conventional concrete [20-22]. Studies incorporating fly ash with recycled concrete aggregates reveal increased water absorption and decreased electrical resistivity [23-25]. RFA concrete, with compressive strengths and modulus of elasticity (MOE) ranging from 70 % to 90 % of NA concrete, sur- passes traditional concrete in terms of CO_2 emissions and cost-effectiveness [24, 25]. Introducing waste carbon

fibers into RFAs further improves mechanical performance [26, 27].

Various techniques such as freeze-thaw, mechanical grinding, ultrasonic treatment, heat treatment, and chemical treatment have been employed to enhance the performance of RAC in reducing adherent mortar [28, 29]. Researchers have explored methods such as adding mineral admixtures, altering mix designs, and adjusting the mixing procedure to improve compressive strength and mechanical properties in aggregates [30-32]. Aggregate replacement techniques such as direct weight replacement, equivalent mortar replacement, and direct volume replacement have been studied [31]. A multi-technique method that includes cement paste removal and fly ash treatment significantly enhances the compressive strength and elastic modulus of RFA concrete [33-35]. The choice of method impacts workability and performance [34, 35]. Despite various techniques applied in material processing, mix design, mixing, and curing stages to enhance RA concrete, the particle packing density (PPD) within a specific cylindrical volume using a conventional compaction technique has not been applied to RAC mix design. This study aims to implement this mixing approach because of its simplicity and effectiveness in producing RAC, aligning with circular economy principles of material reuse. However, ensuring industry-standard quality and performance remains crucial. While previous research involved substituting natural fine aggregates (NFAs) with varying proportions of RFA, a technique to determine the optimal replacement percentage for maximal strength has not been proposed. Our study addresses this gap while also exploring mechanical properties and other characteristics.

Our work's novelty lies in the following aspects:

- It investigates the efficient use of RFA for sustainable concrete production with five distinct RFA levels (0 %, 25 %, 50 %, 75 %, and 100 %) within the concrete mix.
- It examines the influence of various RFA ratios on concrete properties, such as workability, and hardened properties such as compressive strength, flexural strength, splitting tensile strength, and MOE, using established compacting procedures.
- It uses X-ray diffraction (XRD), scanning electron microscopy (SEM), and energy-dispersive X-ray spectroscopy (EDAX) to analyze microstructural attributes and identify significant disparities with the control mixture.
- It examines the effects of RFAs on morphology and composition, encompassing chemical and mineralogical dimensions, to understand their role in generating microstructure. This insight can help formulate sustainable concrete with optimized RFA content.

2. Materials and mixing approach

2.1 Materials

The study utilizes ordinary Portland cement of grade 43, adhering to IS 269-2015 [36] standards. The cement's physical properties reveal a 31 % consistency, an initial setting time of 58 minutes—surpassing the minimum requirement of 30 minutes—and a final setting time of 435 minutes, significantly shorter than the standard's maximum limit of 10 hours. With a specific gravity of 3.11, insights into the density and quality of the cement can be derived. Fine aggregate is sourced from natural sand with a particle size of 4.75 mm and below, meeting IS 383–2016 [37] standards. Natural coarse aggregates (NCAs), conforming to IS 383–2016 [37], consist of crushed stone with sizes ranging from 4.75 mm to 20 mm. These aggregates pass through a 40 mm sieve but are retained on a 4.75 mm sieve. Of these, 90.76 % pass through a 20 mm sieve, and 4.52 % pass through a 10 mm sieve. RFAs, obtained from crushed concrete from CDW, undergo processing with an impact crusher. These RFAs meet size specifications, passing through a 4.75 mm sieve. A chemical admixture in accordance with IS 9103-1999 [38], namely, super-plasticizer (C-MAX), is added at a rate of 1 % by weight of cement. Potable water is utilized for the mixing and curing processes.

2.2 Mixing approach

In the formulation of concrete mixes, PPD is a pivotal tool, facilitating the development of high-performance and sustainable blends [39, 40]. This process entails the careful selection of an optimal aggregate mixture and the assessment of particle size distribution to emulate efficient packing. The objective is to attain maximum packing density and minimize void content, thereby yielding dense and durable concrete structures. Employing PPD not only improves workability but also enhances material efficiency and promotes compatibility with RAs, leading to heightened durability, reduced permeability, and enhanced strength in the resulting concrete.

Table 1 presents data on the weight of various fine aggregate mixtures at specific cylindrical volumes, considering a standard compaction effort. The table illustrates the impact of different ratios of RFA to NFA on the weight or density of the mixes after compaction. Notably, the mixture denoted as RFA 25 + NFA 75 exhibits the highest weight (15.27 units), signifying that this blend, comprising 25 % RFAs and 75 % NFAs, achieves the densest composition under standard compaction conditions. Conversely, the mixture labeled RFA 100 displays the lowest weight (13.27 units), indicating that a composition entirely composed of RFAs attains the lowest density given the

Table 1. Weight in specific cylindrical volume by giving the standard effect of compaction.

S. no	Fine aggregate mixtures	Weight (kg)
1	NFA	14.52
2	RFA 25 + NFA 75	15.27
3	RFA 50 + NFA 50	14.74
4	RFA 75 + NFA 25	13.84
5	RFA 100	13.27

Table 2. Different properties of aggregates.

Property	NFA	RFA	NCA	Standard limits
Bulk density (kg/m^3)	1625	1580	1740	1200–1750
Specific gravity	2.675	2.654	2.754	2.30–2.90
Water absorption (%)	0.51	1.12	0.62	≤ 2.0 (IS 2386 part 3) [41]
Abrasion loss (%)	15.52	18.44	25.43	< 30 (IS 2386 part 4) [42]
Crushing value (%)	16.21	17.34	26.24	< 30 (IS 2386 part 4) [42]
Impact value (%)	15.31	18.74	17.31	< 30 (IS 2386 part 4) [42]

Table 3. Concrete mixture composition (kg/m^3).

S. no.	Mixture ID	NFA	RFA	NCA	Cement	W/C Ratio	Admixture
1	RC	444.48	0	1511	400	0.5	4
2	RFA 25	333.36	111.12	1511	400	0.5	4
3	RFA 50	222.24	222.24	1511	400	0.5	4
4	RFA 75	111.12	333.36	1511	400	0.5	4
5	RFA 100	0	444.48	1511	400	0.5	4

RC – reference concrete NFA – natural fine aggregates NCA – natural coarse aggregates RFA – recycled fine aggregates

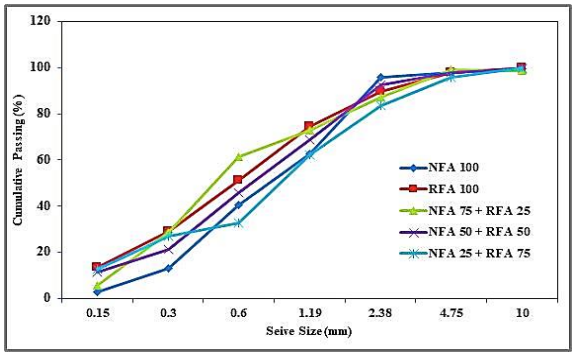


Fig. 1. Fine aggregate gradation curve for various mixtures.

specified compaction conditions. The remaining mixtures, namely, RFA 50 + NFA 50, NFA 100, and RFA 75 + NFA 25, exhibit intermediate weights, arranged in descending order of density. The densest mixture is attained with RFA 25 + NFA 75, and the other mixtures are listed in decreasing order of density. This finding succinctly summarizes the relative densities of the different mixtures. The reference to Fig. 1 suggests the existence of a complementary visual representation, likely in the form of a graph or chart, illustrating the distribution of particle sizes within the coarse aggregates used in the various combinations.

Such a figure would provide additional insights into the physical characteristics of the aggregates and how they influence the density and compaction behavior of the mixtures. The different properties of NFAs, RFAs and NCAs are shown in Table 2.

3. Mix proportions

Five concrete mixtures, with a target strength of 27 MPa, were formulated to examine the viability and influence of RFAs on the mechanical behavior of concrete. The compositions of these mixtures are outlined in Table 3. The reference concrete (RC) comprises NAs, whereas RFA 25, RFA 50, RFA 75, and RFA 100 integrate varying proportions of RFAs (25 %, 50 %, 75 % and 100 %, respectively) in lieu of NFAs. All mixtures maintain a consistent water–cement ratio of 0.50 and adopt the weight batching method [43].

4. Testing methodology

A series of experiments was conducted to comprehensively assess various facets of structural concrete. Key parameters such as compressive strength, flexural strength, splitting tensile strength, and MOE were meticulously measured. In addition, the workability of fresh concrete was evaluated, as depicted in Fig. 2. The microstructural characteristics of concrete samples from different combinations were further scrutinized through advanced techniques, including SEM, EDAX and XRD analyses.

4.1 Workability

The workability of concrete mixtures, incorporating various ratios of recycled CDW as fine aggregates, is evaluated using the IS 1199-1959 [44] standard through a slump test.

4.2 Compressive strength

A set of 30 specimens was generated utilizing steel cube molds with dimensions of 15×15×15 cm. These molds were filled with various concrete mixtures, cast, cured, and then subjected to compressive strength testing. The investigation

(a) (b) (c)

Fig. 2. Testing of various samples: (a) compressive strength cubes; (b) flexural strength beam; (c) splitting tensile strength cylinder.

involved assessing the compressive strength of distinct concrete compositions by utilizing a 2000 KN compression testing machine. Three specimens were tested for each mixture, and the average compressive strength value was computed from these results. The evaluations were performed at intervals of 7 and 28 days after casting.

4.3 Flexural strength

The experimental procedure encompassed the fabrication, curing, and examination of 30 concrete specimens, each shaped in steel rectangular molds measuring 50×10×10 cm. The primary objective of the study was to assess the flexural strength of each concrete blend, gauging its capacity to endure bending or deformation without fracture. With the use of a 2000 KN flexural testing apparatus, three specimens for each mixture were tested, and average flexural strength values were derived from the results. The testing occurred 7 and 28 days after the initial casting.

4.4 Splitting tensile strength

The research entailed the production of 30 concrete specimens utilizing steel cylindrical molds with a diameter of 15 cm and a height of 30 cm. Subsequently, these specimens underwent splitting tensile strength testing to assess the efficacy of diverse concrete mixtures. The tests were performed using a compression testing machine with a 2000 KN capacity, and the splitting tensile strength was determined based on the average value derived from three specimens.

4.5 Modulus of elasticity

This research delved into the nuances of concrete behavior by examining the MOE. Cylindrical specimens measuring 150×300 mm, with a height-to-diameter ratio of 2.0, were utilized for this assessment, following the guidelines specified in IS 516 part 6 of 2006 [45]. These meticulous evaluations yielded essential insights into the structural integrity and deformability of the concrete under various loads.

4.6 X-ray diffraction analysis

XRD is applied to scrutinize the phase characteristics and conduct mineralogical assessments of crystalline minerals present in concrete powder samples, particularly in diverse mixtures incorporating RFAs. The Bruker D-8 advanced diffractometer system is employed to scan the samples across an angular range of 3°–70°, with a scanning speed of 2° per minute and a sampling interval of 0.005° at an angle of 2°. Subsequently, the scans are analyzed using the JADE 7 X-ray diffraction software. The resulting peak intensities are visually represented against 2 theta degrees on the x-axis, spanning from 0° to 70°, with intensity plotted on the y-axis.

4.7 Scanning electron microscopy analysis

SEM analysis was conducted to characterize the materials of RFA, involving a comprehensive examination of the microstructure and surface morphology of concrete samples originating from various concrete mixtures. The analysis was carried out using the JSM 6610V scanning electron microscope at the University Science Instrumentation Centre in Delhi, operating at 30 kV.

5. Results and discussion

5.1 Workability

The slump values outlined in Table 4 indicate a diminishing trend in concrete slump as the proportion of RFAs increases.

This observation suggests that the workability of concrete incorporating RFA falls within a moderate range of 50 mm to 100 mm. This decline is attributed to RFA's higher water absorption capacity compared with NAs, primarily due to the presence of aged adhered mortar on its surface. As the content of RFA rises, the increased water absorption becomes more prominent, leading to a decrease in slump. This pattern aligns with findings from a prior study that noted a slump reduction with increased RCA content in concrete mixtures [46, 47]. The results of the slump test, along with the graphical representation in Fig. 3, provide compelling evidence that the incorporation of RFAs in concrete mixtures contributes to a slump reduction, influencing the concrete's flow and ease of handling.

Table 4. Slump value for various mixtures.

Concrete mix	Slump (mm)
RC	111
RFA 25	104
RFA 50	98
RFA 75	87
RFA 100	76

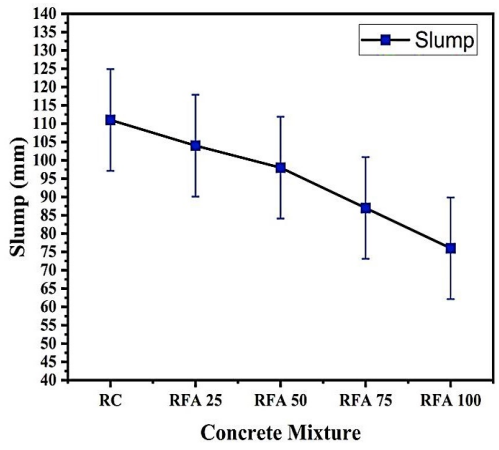


Fig. 3. Slump value for different concrete mixtures.

5.2 Compressive strength

The consolidated data are presented in Table 5, shedding light on how RFAs affect the compressive strength of concrete mixes. The results indicate that, in contrast to the reference mixture, the compressive strength of the concrete diminishes as the proportion of RFA in the mixes increases. Notably, RFA 25 exhibited the highest compressive strength, with a 25 % replacement of fine aggregates being identified as the optimal ratio for RFAs. The enhanced performance of RFA 25 is attributed to improved particle packing and reduced voids. The correlation between strength and the percentage of fine aggregates passing through the 600- to 150-micron range is significant, with a more substantial difference resulting in higher strength within certain limits. Higher replacement percentages of RFA increase void content and hinder strong bonding between aggregates, consequently leading to a reduction in strength. RFA 25 demonstrated the highest compressive strength at 36.71 MPa, while RFA 100 exhibited the lowest compressive strength at 23.01 MPa.

Drawing on the observations of Wagih et al. [48], the introduction of RFAs into concrete mixes led to a reduction in compressive strength of 20 %–34 % at 7 days and 30 %–38 % at 28 days when utilizing 100 % RFA replacement. However, the current study indicates a more moderate decrease in flexural

Table 5. Compressive strength variation in percentage for various combinations.

Concrete mix	Compressive strength (MPa)			
	7 days	% variation concerning RC	28 days	% variation concerning RC
RC	22.71	Reference	31.99	Reference
RFA 25	26.72	+17.65	36.71	+14.75
RFA 50	20.45	−9.95	29.78	−6.90
RFA 75	17.13	−24.57	24.86	−22.28
RFA 100	15.92	−29.89	23.01	−24.94

Note + sign represents an increase in strength and – sign represents a decrease in strength

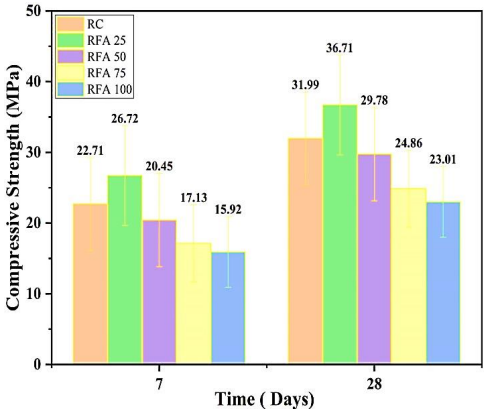


Fig. 4. Compressive strength variations for different concrete mixtures.

Table 6. Flexural strength variation in percentage for different mixtures.

Concrete mix	Flexural strength (MPa)			
	7 days	% variation concerning RC	28 days	% variation concerning RC
RC	3.18	Reference	3.92	Reference
RFA 25	3.41	+7.23	4.16	+4.59
RFA 50	3.12	−1.88	3.89	−0.76
RFA 75	2.89	−9.11	3.75	−4.33
RFA 100	2.60	−18.23	3.48	−11.22

+ Sign represents an increase in strength and – sign represents a decrease in strength

strength, with only a 7 %–10 % reduction after 50 % RFA replacement and a 25 %–29 % decrease with 100 % RFA replacement. These findings underscore the efficacy of the PPD mixing approach through standard compaction, considering weight within a specific cylindrical volume, in preserving flexural strength in concrete mixes. This study aligns with the outcomes of Kessal et al. [49] and lends support to the conclusions drawn by Vintimilla et al. [50]. These earlier studies also emphasized the crucial role of RFAs in achieving desirable concrete strength. Overall, this research contributes to the expanding body of knowledge on optimizing concrete mixtures for enhanced strength characteristics.

5.3 Flexural strength

The detailed outcomes in Table 6 provide a comprehensive understanding of the study. This investigation found that the flexural strength of concrete exhibits variations in terms of time and the amount of RFA replacement. The graphical representation illustrates a consistent pattern of flexural strength diminishing as the percentage of RFA replacement increases, indicating a negative correlation between higher RFA levels and concrete's flexural strength. Notably, after 28 days, RFA 25 demonstrated the highest flexural strength at 4.16 MPa, while RFA 100 exhibited the lowest strength at 3.48 MPa.

This study found a marginal reduction in flexural strength, with only a 1 %–2 % decrease following 50 % RFA replacement and a 10 %–15 % decrease with 100 % RFA replacement. This observation underscores the efficacy of the PPD mixing approach through standard compaction, considering weight within a specific cylindrical volume, in maintaining flexural strength in concrete mixes. This finding diverges from prior research that indicated a 5 %–10 % strength decrease after 50 % RFA replacement and a 15 %–20 % decrease for 100 % replacement [14]. Consistent with earlier research [11, 51], this study establishes a clear association between higher RFA replacement percentages and reduced flexural strength in concrete mixtures, as numerically depicted in Table 6 and visually presented in Fig. 5. These outcomes align with patterns observed in compressive strength studies, reaffirming and contributing to the existing body of knowledge by corroborating similar findings from other researchers.

Table 7. Splitting tensile strength variation in percentage for different mixtures.

Concrete mix	Splitting tensile strength (MPa)			
	7 days	% variation concerning RC	28 days	% variation concerning RC
RC	2.19	Reference	2.51	Reference
RFA 25	2.52	+15.0	2.84	+13.14
RFA 50	2.14	−2.28	2.48	−1.19
RFA 75	1.98	−9.58	2.34	−6.77
RFA 100	1.77	−19.17	2.19	−12.74

Note: + sign represents increase in strength and – sign represents decrease in strength

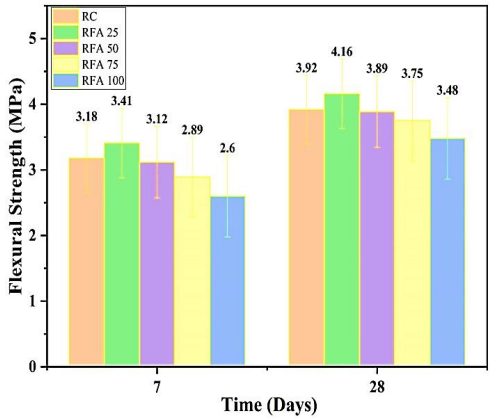


Fig. 5. Flexural strength variations for different concrete mixtures.

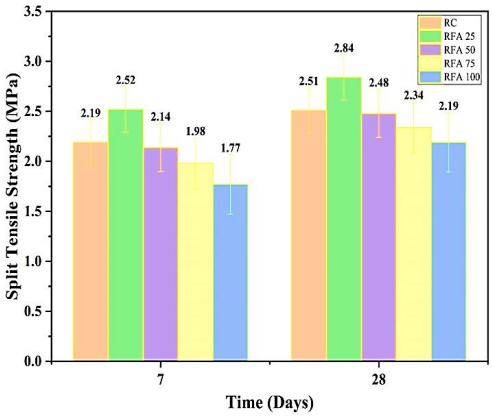


Fig. 6. Splitting tensile strength variations for different concrete mixtures.

5.4 Splitting tensile strength

The findings from these tests are systematically summarized in Table 7. The trends in splitting tensile strength variations over the 7- and 28-day periods are visually elucidated in Fig. 6. This figure effectively portrays the impact of varying percentages of RFA replacement on splitting tensile strength relative to

Table 8. Concrete MOE with different codes.

Concrete mix	Compressive strength (MPa)	Modulus of elasticity (GPa)		
		Experimental (E_c)	As Per IS 456-2000 (E_C)	As Per ACI code (E_C)
RC	31.99	31.42	28.27	26.77
RFA 25	36.71	29.61	30.29	26.68
RFA 50	29.78	24.04	27.28	25.83
RFA 75	24.86	22.68	24.92	23.61
RFA 100	23.01	20.31	23.98	22.72

the reference mixture. The observed trends in splitting tensile strength closely mirror those witnessed in compressive strength variations. Notably, as higher proportions of RFA replacement are employed, the splitting tensile strength of the concrete mixture tends to be lower than that of the reference mixture. Meticulous analysis of the results shows that the concrete mixture labeled "RFA 25" exhibits the highest splitting tensile strength (2.84 MPa) among all the tested mixtures. Conversely, the concrete mixture labeled "RFA 100" demonstrates the lowest splitting tensile strength (2.19 MPa) among the various mixtures examined.

In comparison to this study, where RFA is replaced using the PPD mixing approach through standard compaction in terms of weight within a specific cylindrical volume, only a 12 % reduction occurred at the 100 % replacement level. In contrast, Ashraf M. Wagih's work demonstrated a 24 % decline in splitting tensile strength at a 100 % replacement level [48]. These outcomes align with the findings of Kessal et al., establishing consistency with existing research [49, 50].

5.5 Modulus of elasticity

The MOE serves as a metric to assess the deformation capacity of both conventional concrete and RAC. MOE measures how much a material deforms when subjected to a load and subsequently returns to its original shape upon load release. The primary objective of this study is to ascertain the influence of RFA content on the deformation characteristics of concrete. The MOE at 28 days demonstrates a decline as the RFA content in the concrete increases. Table 8 delineates that RFA's brittleness and water absorption significantly contribute to this reduction, leading to more pronounced deformation in structural components made with RFA compared with those using NA. Consequently, an increase in RFA content correlates with a reduction in the MOE in concrete.

Fig. 7 visually presents the decline in the MOE with increasing RFA content, despite the potential benefits of the particle density method in enhancing strength and hydration. The comparison between compressive strength and the MOE of plain concrete for different RFA substitution percentages reinforces this observed pattern. The findings also suggest that various factors, including aggregate volume and stiffness, binder phase stiffness, and characteristics of the interfacial transition zone between aggregate and paste, contribute to the decrease in

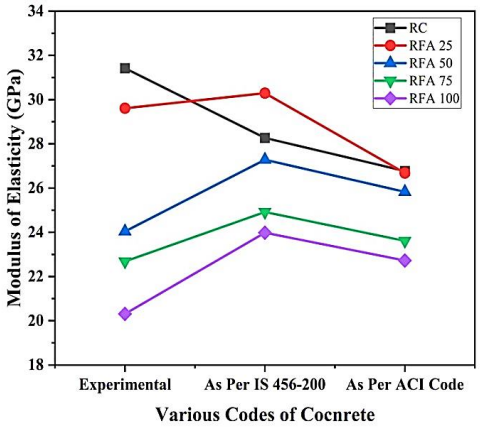


Fig. 7. Modulus of elasticity variation with different codes.

MOE with the increasing amount of RFA replacement. These results align with earlier studies that also identified a similar pattern of decreasing MOE as the quantity of RCA in concrete increased [48, 52].

5.6 X-ray diffraction analysis

XRD is a robust technique for elucidating the phase characteristics and mineralogical composition of crystalline minerals in a concrete sample. This method is particularly valuable when examining the effects of diverse mixtures that integrate RFAs. The angles at which diffraction peaks appear for different compounds in the concrete mixtures are recorded in Table 9. This table provides the degree values for calcium silicate hydroxide (CSH), ettringite, and calcium hydroxide (CH) for each concrete mixture, offering in-depth insight into the mineralogical aspects of the specimens by elucidating the diffraction peak angles for various compounds. The interpretation and discussion noted that the angles for CSH exhibit minimal variation, suggesting a consistent presence of this compound. The slight decrease in angles for RFA 25 indicates a subtle alteration in CSH composition compared with the RC. Ettringite angles show some variation, with RFA 25 displaying a lower angle, suggesting potential modifications in the crystalline structure. Increasing angles with higher RFA content hint at changes in the formation or orientation of ettringite crystals.

CH angles undergo more pronounced changes, notably in RFA 75 and RFA 100, where a substantial increase is ob-

Table 9. Diffraction peak angles of various compounds for different concrete mixtures.

Concrete mix	CSH (degree)	Ettringite (degree)	CH (degree)
RC	26.450	36.780	47.299
RFA 25	26.337	34.146	50.175
RFA 50	26.596	36.606	50.110
RFA 75	26.575	39.385	60.001
RFA 100	26.571	39.523	59.930

Note: CSH – calcium silicate hydroxide CH – calcium hydroxide ettringite – hydrated calcium aluminum sulfate hydroxide

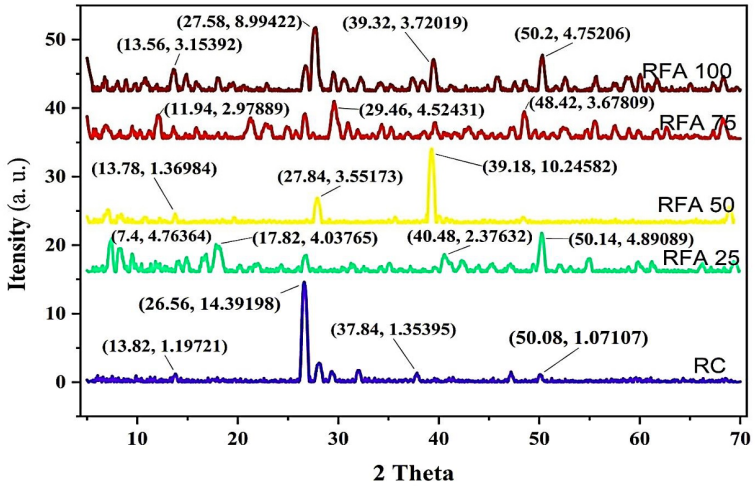


Fig. 8. XRD pattern for different concrete mixtures.

served, and potentially indicating variations in hydration products due to the presence of RFAs. Furthermore, the XRD results demonstrate the influence of RFAs on the phase composition of minerals within concrete. Specifically, the net intensity of peaks corresponding to CSH, CH, and ettringite decreases as the percentage of RFA replacement increases. This finding suggests a diminishing density of total CSH in the concrete with greater incorporation of RFAs. Fig. 8 presents the XRD patterns for various concrete mixtures, including RC, RFA 25, RFA 50, RFA 75, and RFA 100, offering a graphical depiction of how peak intensities and positions vary across these mixtures. These findings contribute to a comprehensive understanding of the effect of RFA on the mineralogical characteristics and overall composition of concrete mixtures.

As revealed by XRD studies, the addition of RFAs induces subtle changes in the phase composition of CSH, CH, and ettringite within concrete. The recorded data exhibit peaks at approximately 26.5°, 36°, and 50° for 2-theta angles, signifying the presence of various compounds such as CSH, ettringite, and CH across all combinations. While the strength of these peaks suggests the substantial presence of these substances, an important detail to note is that the concrete samples utilized in the investigation lacked microscopic homogeneity. This observation aligns with Silva et al.'s [53] study, which examined cement paste containing 12 % concrete floor polishing waste, highlighting comparable findings between the two studies [14]. In summary, XRD results imply that the incorporation of RFAs introduces subtle changes in the composition and crystalline structure of specific compounds, emphasizing the influence of RFA on the hydration process and overall mineralogical characteristics of the concrete. These nuanced insights are pivotal for evaluating the long-term durability and performance of concrete mixtures integrating RFAs.

5.7 Scanning electron microscopy analysis

In this investigation, SEM analysis is employed to scrutinize the characteristics of RFA concrete. Fig. 9 provides micrographs obtained after 28 days for distinct concrete mixtures that underwent SEM analysis.

These micrographs offer valuable insights into the microstructure of different concrete mixtures, revealing the formation of crucial hydration products at this stage. These hydration products, including CH, CSH, and ettringite, play a pivotal role in determining concrete strength. The micrographs illustrate distinctive characteristics: hexagonal crystal shapes correspond to CH, flower-like structures represent CSH gel, and needle-like structures denote ettringite. The test results reveal significant insights into the impact of different proportions of RFA on concrete microstructure and strength. The introduction of 25 % RFA into the concrete mixture enhances its microstructure, fostering the development of a dense cement paste. This

improvement strengthens the bond between aggregates, consequently elevating the compressive strength of the concrete. Conversely, the inclusion of 50 % RFA results in less dense concrete compared with the reference mix, thereby reducing both density and compressive strength. At 75 % RFA replacement, the concrete structure exhibits voids and loose arrangements, further diminishing the compressive strength. A complete 100 % RFA replacement yields a porous microstructure, where the cement surrounding the RFAs undergoes incomplete hydration, resulting in fewer hydration products. SEM analysis elucidates that lower RFA mixture proportions generate a more substantial amount of CSH gel, contributing to a denser cement paste matrix and increased concrete strength. However, excessive RFA proportions weaken the mixture's structural integrity due to inadequate CSH production, resulting in reduced strength. These findings align with a previous study [28, 54], reinforcing the consistency of the results with existing research. In summary, the SEM analysis in this study provides valuable insights into how varying RFA proportions influence concrete microstructure and strength. The micrographs in Fig. 9 highlight distinct features corresponding to different hydration products, emphasizing that optimal RFA replacement proportions can enhance concrete properties, while excessive replacement can lead to weaker structures. These insights bear significance for the design of sustainable and high-performance concrete mixtures.

6. Cost–benefit analysis

Table 10 presents a comprehensive overview of the cost–benefit analysis conducted in this study, comparing the scheduled rate of the Road and Building Department Delhi Division, Government of Delhi with the prevailing market rates for various construction materials. The materials considered in this

Table 10. Cost of mixtures.

Elements	Unit cost per kg	RC (kg/m^3)	RFA 25 (kg/m^3)	RFA 50 (kg/m^3)	RFA 75 (kg/m^3)	RFA 100 (kg/m^3)
Cement	6.6	400	400	400	400	400
Admixture	0.9	4	4	4	4	4
NFA	0.75	444.48	333.36	222.24	111.12	0
NCA	0.65	1511	1511	1511	1511	1511
RFA	0.40	0	111.12	222.24	333.36	444.48
Total (INR)		3959.11	3920.21	3881.32	3842.43	3803.54
% variation concerned with RC		-	0.98	1.96	2.94	3.96

RC – reference concrete NFA – natural fine aggregates NCA – natural coarse aggregates RFA – recycled fine aggregates

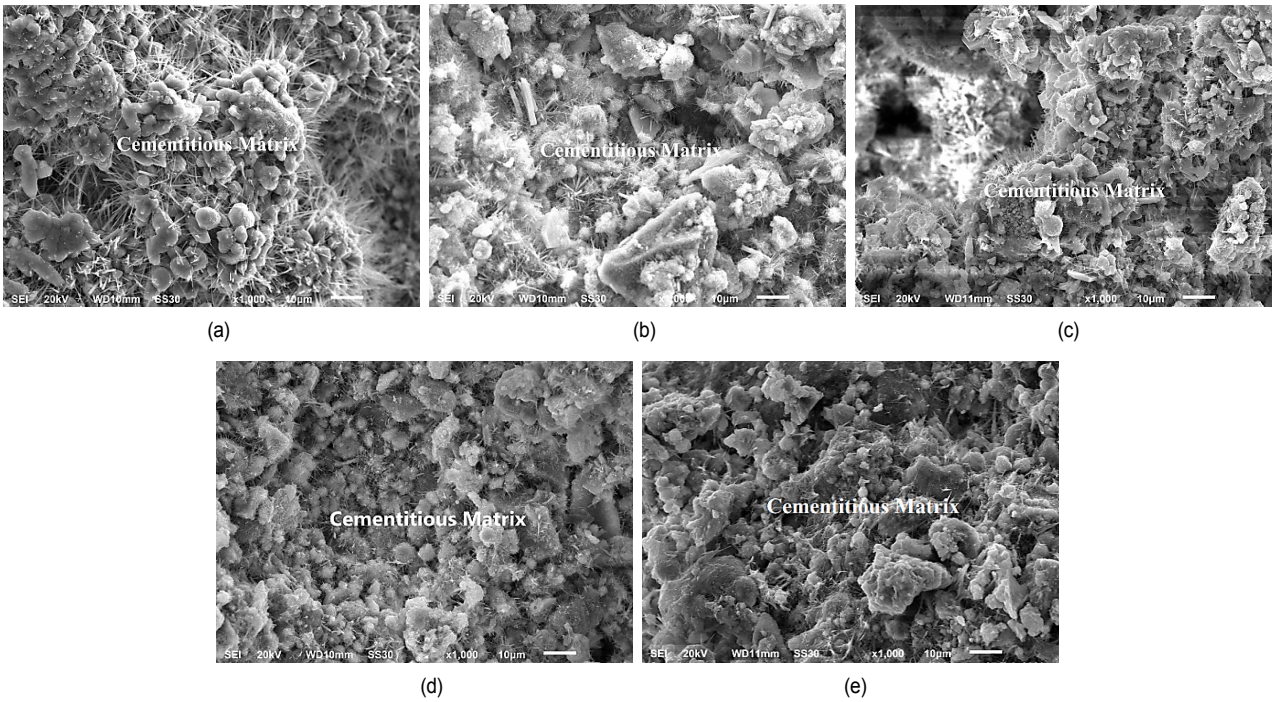


Fig. 9. Different concrete mixtures' SEM micrographs: (a) RC; (b) RFA 25; (c) RFA 50; (d) RFA 75; (e) RFA 100.

analysis include:

- Cement priced at INR 6.6 per kilogram.
- NFAs priced at INR 0.75 per kilogram.
- NCAs priced at INR 0.65 per kilogram.
- RFAs priced at INR 0.40 per kilogram.

RFAs are produced utilizing an impact crusher at the Delhi Metro Rail Corporation's IL&FS CDW Recycling Plant in Delhi, India, using discarded concrete.

The objective of this cost analysis is to quantify potential savings achievable through the utilization of CDW in comparison to NAs. Table 10 illustrates the material costs and the resultant expenses incurred by different mixtures. As depicted in the table, substituting NAs with RFAs can lead to cost savings ranging from 1 % to 4 %, depending on the quantity of replaced RFA. This approach not only offers cost savings but also delivers concrete of comparable quality to the control, making it a favorable choice from both economic and performance perspectives [55].

7. Conclusions

This paper investigates RFAs from CDW in sustainable concrete. The findings highlight reduced environmental impact without compromising structural integrity, emphasizing controlled particle size distribution. From the study, the following conclusions may be drawn:

1) The mechanical properties of concrete exhibit an optimum replacement percentage for RFAs at 25 %. Concrete achieves superior performance at this level, with a decline in strength characteristics observed at 100 % RFA replacement. Careful evaluation of the RFA replacement proportion is essential for ensuring intended concrete mixture performance.
2) A recommended replacement of 25 % RFAs in concrete yields higher strength compared with NFAs. RFA 25 demonstrates a 14.75 % increase in compressive strength, a 6.61 % increase in flexural strength, and a 13.14 % increase in splitting tensile strength compared with RC. The strength of 50 % RFA is comparable to that of 100 % NFAs.
3) The pozzolanic reaction enhances strength at lower replacement percentages by forming more hydration products, while higher levels result in decreased strength due to incomplete hydration. Optimal replacement percentage determination relies on microstructure and chemical composition analysis to achieve maximum strength.
4) The experimental MOE for concrete, incorporating both NA and RA, is lower than the theoretical value due to flaws, irregularities in recycled aggregates, and variations in composition.
5) RAs significantly reduce environmental impact by curbing natural resource demand, conserving resources, and minimizing carbon footprint. Their use promotes sustainable building methods, reduces landfill waste, and encourages resource reuse, resulting in a 10 % cost–benefit ratio and less negative environmental effects in construction projects.

Incorporating RFAs in concrete is a promising sustainable practice, with marginal mechanical property reduction. Microstructural analysis offers relevant insights. Further research is needed for optimal mix designs, long-term performance, and economic feasibility, marking a positive step in reducing construction's environmental impact.

Acknowledgments

The authors acknowledge the staff of the Department of Civil Engineering at Delhi Technological University, Shahbad Daultpur Village, Rohini, New Delhi, for their significant support and help during the research.

References

[1] O. Zaid, S. R. Zamir Hashmi, F. Aslam and H. Alabduljabbar, Experimental study on mechanical performance of recycled fine aggregate concrete reinforced with discarded carbon fibers, *Front. Mater.*, 8 (11) (2021) 1-12.

[2] D. Hoornweg and P. Bhada-Tata, A global review of solid waste management - review, global management, solid waste, *World Bank Urban Dev. Ser. Knowl. Pap.*, 1 (11) (2012) 1-116.

[3] O. E. Gjørv, Durability of concrete structures, *Arab. J. Sci. Eng.*, 36 (2) (2011) 151-172.

[4] T. R. Sonawane and S. S. Pimplikar, Use of recycled aggregate concrete, *IOSR J. Mech. Civ. Eng.* (2012) 52-59.

[5] S. R. Boyle, Evaluation of recycled concrete for use as aggregates in new portland cement concrete pavements, *Master's Thesis*, Washington State University, USA (2013).

[6] W. Z. Taffese, Suitability investigation of recycled concrete aggregates for concrete production: an experimental case study, *Adv. Civ. Eng.* (2018) 8368351.

[7] S. Bansal and S. K. Singh, A sustainable approach towards the construction and demolition waste, *Int. J. Innov. Res. Sci. Eng. Technol.*, 3 (2) (2013) 9226-9235.

[8] S. Bansal, S. K. Singh and J. Kurian, Construction and demolition (C&D) waste recycling in new delhi, *4th Int. Fib Congr. 2014 Improv. Perform. Concr. Struct. FIB 2014 - Proc.*, 6 (2014) 286-289.

[9] S. Bansal and S. K. Singh, Sustainable handling of construction and demolition (C & D) waste, *Int. J. Sustain. Energy Environ. Res.*, 4 (2) (2015) 22-48.

[10] Gagan and S. Arora, Recycled Aggregates: a sustainable solution of construction and demolished waste, *J. Mech. Civ. Eng.* (2015) 58-63.

[11] M. Surya, K. Rao VVL and P. Lakshmy, Recycled aggregate concrete for transportation infrastructure, *Procedia - Soc. Behav. Sci.*, 104 (11) (2013) 1158-1167.

[12] L. Evangelista and J. De Brito, Concrete with fine recycled aggregates : a review, *European Journal of Environmental and Civil Engineering*, 18 (2) (2014) 129-172.

[13] J. Yu, D. Liu and Z. Zhang, Durability and life prediction

analysis of recycled aggregate concrete with ceramic waste powder under freeze-thaw conditions based on impact-echo method and grey-markov model, *Front. Mater.*, 9 (10) (2022) 1-13.

[14] H. S. Joseph, T. Pachiappan, S. Avudaiappan and E. I. S. Flores, A study on mechanical and microstructural characteristics of concrete using recycled aggregate, *Materials (Basel).*, 15 (21) (2022) 15217535.

[15] L. Courard, M. Rondeux, Z. Zhao and F. Michel, Use of recycled fine aggregates from c & dw for unbound road sub-base, *Materials (Basel)*, 13 (2020) 2994.

[16] C. Shi, Y. Li, J. Zhang, W. Li, L. Chong and Z. Xie, Performance enhancement of recycled concrete aggregate - a review, *J. Clean. Prod.*, 112 (2016) 466-472.

[17] H. Yaprak, H. Y. Aruntas, I. Demir and O. Simsek, Effects of the fine recycled concrete aggregates on the concrete properties, *International Journal of Physical Sciences*, 6 (10) (2011) 2455-2461.

[18] A. Kumar and G. J. Singh, Improving the physical and mechanical properties of recycled concrete aggregate: A state-of-the-art review, *Eng. Res. Express*, 5 (1) (2023) 012007.

[19] D. D. Nguyen, Evaluating the possibility of replacing natural fine aggregates in concrete with recycled aggregates, *Eng. Technol. Appl. Sci. Res.*, 11 (6) (2021) 7805-7808.

[20] K. Naouaoui and T. Cherradi, A case study on the mechanical and durability properties of a concrete using recycled aggregates, *Civ. Eng. J.*, 7 (11) (2021) 1909-1917.

[21] K. Osterminski, R. B. Polder and P. Schießl, Long term behaviour of the resistivity of concrete, *Heron*, 57 (3) (2012) 211-230.

[22] H. Layssi, P. Ghods, A. R. Alizadeh and M. Salehi, Electrical resistivity of concrete, *Concr. Int.*, 2 (2015) 293–302.

[23] H. Oleiwi, Y. Wang, N. Xiang, L. Augusthus-Nelson, X. Chen and I. Shabalin, An experimental study of concrete resistivity and the effects of electrode configuration and current frequency on measurement, *6th Int. Conf. Durab. Concr. Struct. ICDCS 2018*, University of Leeds, UK., 6 (2018) 592-599.

[24] M. Quattrone, S. C. Angulo and V. M. John, Energy and CO_2 from high performance recycled aggregate production, *Resour. Conserv. Recycl.*, 90 (2014) 21-33.

[25] B. J. Zhan, D. X. Xuan and C. S. Poon, Enhancement of recycled aggregate properties by accelerated CO_2 curing coupled with limewater soaking process, *Cem. Concr. Compos.*, 89 (2018) 230-237.

[26] C. M. M. de A. e. Silva, M. M. L. Pereira, V. M. S. Capuzzo and J. de Brito, Concrete produced with recycled concrete aggregate exposed to treatment methods, *Case Stud. Constr. Mater.*, 18 (11) (2022) 01938.

[27] C. Cakiroglu and G. Bekdaş, Predictive modeling of recycled aggregate concrete beam shear strength using explainable ensemble learning methods, *Sustainability*, 15 (6) (2023) 15064957.

[28] V. W. Y. Tam, X. F. Gao and C. M. Tam, Microstructural analysis of recycled aggregate concrete produced from two-stage mixing approach, *Cem. Concr. Res.*, 35 (6) (2005) 1195-1203.

[29] J. paul Ntitanguranwa, J. Kabano and N. Gasingwa, Reuse of construction and demolished concrete waste by producing affordable high strength concrete block, *Rwanda J. Eng. Sci. Technol. Environ.*, 1 (1) (2018) 14.

[30] M. Ju, J. Jeong, M. Palou and K. Park, Mechanical behavior of fine recycled concrete aggregate concrete with the mineral admixtures, *Materials*., 13 (10) (2020) 1-15.

[31] S. Pradhan, S. Kumar and S. V Barai, Recycled aggregate concrete: particle packing method (PPM) of mix design approach, *Constr. Build. Mater.*, 152 (2017) 269-284.

[32] S. Ismail and M. Ramli, Engineering properties of treated recycled concrete aggregate (RCA) for structural applications, *Constr. Build. Mater.*, 44 (2013) 464-476.

[33] M. K. Rao and D. C. N. S. Kumar, Durability assessment of concrete with class-F fly ash by chloride ion permeability, *Int. J. Recent Technol. Eng.*, 8 (4) (2019) 8831-8836.

[34] G. L. Golewski, Combined effect of coal fly ash (CFA) and nanosilica (nS) on the strength parameters and microstructural properties of eco-friendly concrete, *Energies*, 16 (1) (2023) 452.

[35] B. Uzbaş and A. C. Aydin, Analysis of fly ash concrete with scanning electron microscopy and X-Ray diffraction, *Adv. Sci. Technol. Res. J.*, 13 (4) (2019) 100-110.

[36] ISO 269:2015, *Indian Standard for Ordinary Portland Cement*, Bureau of Indian Standards, India (2017).

[37] ISO 383:2016, *Specification for Coarse and Fine Aggregate for Concrete*, Bureau of Indian Standards, India (2016).

[38] ISO 9103:1999, *Specification for Concrete Admixtures*, Bureau of Indian Standards, India (1999).

[39] S. Fennis and J. Walraven, Using particle packing technology for sustainable concrete mixture design, *Heron*, 57 (2) (2012) 72-102.

[40] B. Demolition, W. Bdw and S. Solution, Particle packing approach for proportioning recycled aggregate from particle packing approach for proportioning recycled aggregate from building demolition waste (BDW) — a sustainable solution, *J. Inst. Eng. Ser. A*, 11 (2022) 686-688.

[41] ISO 2386:1963 Part III, *Methods of Test for Aggregates for Concrete*, Bureau of Indian Standards, India (2021).

[42] ISO 2386:1963 Part IV, Methods of Test for Aggregates for Concrete, Part 4: Mechanical Properties, Bureau of Indian Standards, India (2016).

[43] ISO 10262:2009, *Guidelines for Concrete Mix Design Proportioning*, Bureau of Indian Standards, India (2009).

[44] ISO 1199:2018 Part 1, *Methods of Sampling, Testing, and Analysis for Fresh Concrete*, Bureau of Indian Standards, India (2018).

[45] ISO 516:2020 Part 6, Determination of Drying Shrinkage and Moisture Movement of Concrete Samples, Bureau of Indian Standards, India (2020).

[46] Z. Zhao, S. Remond, D. Damidot and W. Xu, Influence of fine recycled concrete aggregates on the properties of mortars, *Constr. Build. Mater.*, 81 (2015) 179-186.

[47] S. K. Kirthika, S. K. Singh and A. Chourasia, Performance of recycled fine-aggregate concrete using novel mix-proportioning method, *J. Mater. Civ. Eng.*, 32 (8) (2020) 1-13.

[48] A. M. Wagih, H. Z. El-Karmoty, M. Ebid and S. H. Okba, Re-

cycled construction and demolition concrete waste as aggregate for structural concrete, *HBRC J.*, 9 (3) (2013) 193-200.
[49] O. Kessal, L. Belagraa, A. Noui and N. Maafi, Performance study of eco-concrete based on waste demolition as recycled aggregates, *Mater. Int.*, 2 (2) (2020) 123-130.
[50] C. Vintimilla and M. Etxeberria, Limiting the maximum fine and coarse recycled aggregates-type a used in structural concrete, *Constr. Build. Mater.*, 380 (3) (2023) 131273.
[51] E. O. Aiyewalehinmi and T. E. Adeoye, Recycling of concrete waste material from construction demolition, *Am. J. Eng. Res.*, 5 (4) (2016) 182-191.
[52] M. Velay-Lizancos, I. Martinez-Lage, M. Azenha, J. Granja and P. Vazquez-Burgo, Concrete with fine and coarse recycled aggregates: E-modulus evolution, compressive strength and non-destructive testing at early ages, *Constr. Build. Mater.*, 193 (2018) 323-331.
[53] L. A. Silva, B. O. Nahime, E. C. Lima, J. L. Akasaki and I. C. Reis, XRD investigation of cement pastes incorporating concrete floor polishing waste, *Ceramica*, 66 (380) (2020) 373-380.
[54] H. S. Wong et al., Microscopy techniques for determining water–cement (w/c) ratio in hardened concrete: a round-robin assessment, *Mater. Struct. Constr.*, 53 (2) (2020) 1458.
[55] A. Al-Mansour, C. L. Chow, L. Feo, R. Penna and D. Lau, Green concrete: by-products utilization and advanced approaches, *Sustain.*, 11 (19) (2019) 1-30.

Harish Panghal is currently pursuing a Ph.D. in the Department of Civil and Environmental Engineering at Delhi Technological University, Delhi. He holds a bachelor's degree in civil engineering (2013) and a master's degree in structural engineering (2016). His research focuses on addressing the challenges associated with the development of sustainable concrete using construction and demolished waste, exploring potential solutions in this realm.

Awadhesh Kumar holds a master's degree in structural engineering and obtained his Ph.D. in civil engineering in 2006 from the Indian Institute of Technology, Roorkee. Presently, he serves as a Professor in the Department of Civil and Environmental Engineering at Delhi Technological University, Delhi. With a wealth of experience in civil engineering, his expertise extends to advanced concrete techniques and the structural design of tall buildings.

The Open Biomedical Engineering Journal
DOI: 10.2174/0118741207289680240510045617, 2024, 18, e18741207289680
ISSN: 1874-1207
1

RESEARCH ARTICLE **OPEN ACCESS**

Exploring Biomedical Named Entity Recognition *via* SciSpaCy and BioBERT Models

Aman Jolly[1], Vikas Pandey[2], Indrasen Singh[3,*] and Neha Sharma[4]

[1]*KIET Group of Institutions, UP, India*
[2]*Department EE, School of Engineering, BBD University, Lucknow, UP, India*
[3]*School of Electronics Engineering, VIT Vellore, Tamil Nadu, India*
[4]*Delhi Technological University, Delhi, India*

Abstract:

Introduction: Biological Named Entity Recognition (BioNER) is a crucial preprocessing step for Bio-AI analysis.

Methods: Our paper explores the field of Biomedical Named Entity Recognition (BioNER) by closely analysing two advanced models, SciSpaCy and BioBERT. We have made two distinct contributions: Initially, we thoroughly train these models using a wide range of biological datasets, allowing for a methodical assessment of their performance in many areas. We offer detailed evaluations using important parameters like F1 scores and processing speed to provide precise insights into the effectiveness of BioNER activities.

Results: Furthermore, our study provides significant recommendations for choosing tools that are customised to meet unique BioNER needs, thereby enhancing the efficiency of Named Entity Recognition in the field of biomedical research. Our work focuses on tackling the complex challenges involved in BioNER and enhancing our understanding of model performance.

Conclusion: The goal of this research is to drive progress in this important field and enable more effective use of advanced data analysis tools for extracting valuable insights from biomedical literature.

Keywords: Biomedical, BioNER, SciSpaCy, BioBERT, Natural language processing, Named entity recognition.

© 2024 The Author(s). Published by Bentham Open.
This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International Public License (CC-BY 4.0), a copy of which is available at: https://creativecommons.org/licenses/by/4.0/legalcode. This license permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

*Address correspondence to this author at the School of Electronics Engineering, VIT Vellore, Tamil Nadu, India;
E-mail: erindrasen@gmail.com

Cite as: Jolly A, Pandey V, Singh I, Sharma N. Exploring Biomedical Named Entity Recognition *via* SciSpaCy and BioBERT Models. Open Biomed Eng J, 2024; 18: e18741207289680. http://dx.doi.org/10.2174/0118741207289680240510045617

Received: November 14, 2023
Revised: March 13, 2024
Accepted: April 08, 2024
Published: June 05, 2024

Send Orders for Reprints to
reprints@benthamscience.net

1. INTRODUCTION

Named Entity Recognition (NER) [1] is a crucial part of Natural Language Processing (NLP) that involves identifying and categorising named items in written text. Named entities encompass several categories such as personal names, organisational names, and geographical places, among others. The primary objective is to do this by transforming unstructured textual content into a machine-readable format. Named Entity Recognition (NER) is a crucial part of Information Extraction (IE) that involves the detection and classification of named items in unstructured text data. Named entities refer to a broad variety of entities, such as persons, businesses, geographical locations, precise dates, numerical values, and other pertinent categories. Named Entity Recognition (NER) is a prominent area of research that has attracted substantial academic attention and has several practical applications in the field of natural language processing. The use of NER approaches varies across different domains, giving rise to distinct difficulties and possibilities

within this subject. Deep learning techniques for NER have demonstrated encouraging outcomes across diverse areas such as healthcare, legal, and chemical fields. There have been proposals to utilise active learning methodologies, specifically Label Transfer Learning (LTP), as a means to mitigate the expenses associated with data annotation in the field of NER. In addition, there are shared projects and benchmarks, such as the ROCLING-2022 shared task focused on Chinese healthcare NER, which furnish datasets and assessment criteria to facilitate the comparison of diverse NER methodologies. Named Entity Recognition is a specific component within the broader field of Information Extraction that falls within the domain of NLP. The primary objective of this system is to discern and categorise designated items, such as individual names, organisational names, geographical places, dates, and other specific nouns, inside the text that lacks a predetermined framework [2-4]. Named Entity Recognition is an essential preprocessing task in several downstream NLP applications, including but not limited to question answering, conversation systems, and knowledge graph generation [3].

The use of Named Entity Recognition algorithms exhibits variability across many domains, owing to the distinctive characteristics inherent to each specific area [4]. Nevertheless, the procedure often encompasses other essential NLP stages, such as tokenization, part-of-speech tagging, parsing, and model construction [4]. In the field of medicine, for instance, Named Entity Recognition encompasses the task of recognising medical terminology, properties such as negation and severity, and establishing connections between terms and ideas inside domain-specific ontologies [5–9].

In the early stages of development, Named Entity Recognition (NER) systems mostly relied on manually constructed rules, lexicons, and ontologies. However, these systems need extensive dictionaries and meticulous manual feature creation, which constrained their ability to expand and reach maximum performance [10, 11]. In the era of Machine Learning, there are solutions that include machine learning concepts as a potential alternative [12, 13]. However, even these solutions required extensive feature engineering work, limiting their flexibility and effectiveness. The scenario saw significant shifts due to the emergence of the Deep Learning revolution. Recent advancements have brought about advanced deep learning structures like Bi-directional Long Short-Term Memory Networks (BiLSTM) combined with Conditional Random Fields (CRF) and GRAM-CNN [14]. The developments have far beyond the capability of older approaches, demonstrating improved precision and resilience. The greatest significant advancement in Named Entity Recognition (NER) technology occurred with the development of the BERT architecture (Bi-directional Encoder Representations from Transformers). BERT has introduced a new age by using contextual information, allowing NER systems to attain exceptional levels of accuracy and recall [15-18]. The capacity to understand subtle contextual signals has transformed NER accuracy, representing a significant advancement in biomedical text comprehension and data retrieval.

Named Entity Recognition can provide difficulties, particularly in situations involving low-resource languages or topics that necessitate a time-consuming and knowledge-intensive annotation procedure [4, 5, 19]. In order to tackle these issues, scholars have undertaken investigations into many methodologies, including few-shot learning. This particular methodology endeavours to train machine learning models with a very restricted amount of accessible data [19-21]. Furthermore, the integration of entity definition information and domain-specific resources, such as medical dictionaries and ontologies [8, 13, 14, 22], has been shown to enhance the efficacy of Named Entity Recognition models within certain domains [18, 23-32]. Multiple strategies are employed in order to enhance the precision of NER. Several approaches can be identified, including:

1. The process of data augmentation encompasses the generation of supplementary training instances through the modification or synthesis of pre-existing examples. This approach has the potential to enhance the generalisation capabilities of the Named Entity Recognition paradigm [33-36].

2. The hybrid approach involves integrating rule-based methodologies with machine learning techniques, namely Conditional Random Fields (CRF), to enhance the precision of Named Entity Recognition [34, 37, 38].

3. The inclusion of domain-specific resources, such as medical dictionaries and ontologies, together with entity definition information, has been found to enhance the effectiveness of named entity recognition models in certain domains [39-41].

4. Semi-supervised learning is a technique that involves training a model using a little amount of labelled data together with a large amount of unlabeled data. This technique has the potential to improve the effectiveness of named entity recognition models in situations typified by restricted resources, such as low-resource languages or domains where the annotation process requires substantial time, knowledge, and expertise [33, 42, 43]. One way to address the problem of noisy data and the introduction of unwanted patterns during training is to use an iterative training process and data-generating tools. This entails retraining the model using a subset of the original annotated dataset [33, 44, 45].

5. Fine-tuning and transfer learning are often employed techniques in the field of named entity recognition to enhance the efficiency and/or accuracy of NER models [46-48]. Researchers have achieved enhanced precision in NER models across several domains, such as healthcare, legal, and chemical domains, through the use of these methodologies [44, 49]. The job of Named Entity Recognition poses significant challenges in the context of clinical writing, primarily due to the intricate and diverse nature of medical language. Nevertheless, many methodologies may be employed to enhance the precision of Named Entity Recognition in clinical discourse [47].

Several approaches can be identified, including the integration of domain-specific resources, including entity definition information and specialised resources like medical dictionaries and ontologies, which have been shown to enhance the efficacy of named entity recognition models within the medical domain [18, 50]. The hybrid approach involves integrating rule-based methodologies with machine learning techniques, namely CRF, in order to enhance the precision of Named Entity Recognition [51]. Data augmentation is a methodology that entails the creation of supplementary training instances through the alteration or synthesis of preexisting examples. This approach has the potential to enhance the generalisation capabilities of the Named Entity Recognition paradigm [9]. Semi-supervised learning is a methodology that entails training a model using a small quantity of labelled data with a substantial quantity of unlabelled data. This approach has the potential to enhance the efficacy of Named Entity Recognition models in scenarios characterised by limited resources, such as low-resource languages or domains where the annotation process necessitates substantial effort, skill, and domain knowledge [50, 52]. Fine-tuning and transfer learning techniques can be employed to enhance the efficiency and/or accuracy of Named Entity Recognition models [53, 54]. The utilisation of a multilevel named entity recognition framework can effectively tackle the difficulties associated with clinical NER. This framework enables the construction of models that are capable of handling more intricate NER tasks [53, 55]. Through the use of these methodologies, scholars have successfully enhanced the precision of Named Entity Recognition models within the medical field. As an illustration, a research endeavour pertaining to the identification and extraction of adverse drug events from clinical literature by the use of medicine and associated techniques yielded F1 scores of 93.45% for Named Entity Recognition [50]. A recent investigation pertaining to the identification of medication names and their corresponding properties from discharge summaries demonstrated a notable F1-score of 0.91 in the context of the 2018 National NLP Clinical Challenges (n2c2) shared task [50]. Identifying uncommon medical conditions and their associated clinical presentations, together with the influence of socio-economic determinants of health, can provide challenges within the clinical literature. Moreover, the intricate and diverse nature of medical vocabulary presents a formidable obstacle for Named Entity Recognition when applied to clinical material. The concept of stupor poses a challenge in terms of its specific clinical definition [5]. Nevertheless, scholars have put forth various methodologies to enhance the precision of Named Entity Recognition in clinical text [56-59]. These methodologies encompass the integration of domain-specific resources, the adoption of a hybrid approach, data augmentation techniques, semi-supervised learning methods, fine-tuning and transfer learning strategies, as well as the utilisation of a multilevel NER framework [58, 60–62]. Identifying abbreviations and acronyms in clinical material is difficult for Named Entity Recognition algorithms. Nevertheless, many methodologies may be employed to enhance the precision of Named Entity Recognition in clinical discourse containing abbreviations and acronyms. Several approaches can be identified, including the use of domain-specific resources, such as medical dictionaries and ontologies, that enhance the ability of named entity recognition models to identify abbreviations and acronyms in the clinical text [37]. The hybrid approach involves integrating rule-based methodologies with machine learning techniques, namely CRF, in order to enhance the precision of Named Entity Recognition [51]. Data augmentation is a methodology that entails the creation of supplementary training instan-ces through the alteration or synthesis of preexisting examples. This approach has the potential to enhance the generalisation capabilities of the Named Entity Recog-nition model [55]. Fine-tuning and transfer learning are two techniques that may be employed to enhance the efficiency and/or accuracy of Named Entity Recognition models [53].The utilisation of a multilevel named entity recognition framework presents a viable solution to tackle the obstacles encountered in clinical NER. This framework enables the construction of models that cater to more intricate NER tasks [63].

Through the use of these methodologies, scholars have successfully enhanced the precision of Named Entity Recognition models when applied to clinical material containing abbreviations and acronyms. As an illustration, a research investigation on clinical named entity identification attained an F1-score of 0.91 on the 2018 National NLP Clinical Challenges (n2c2) shared task [27, 50]. When applied to the field of biomedical research, the Biomedical Named Entity Recognition (BioNER) technique is an essential part of the process of biomedical text mining [44, 64-66]. The enormous task of recognising and categorising the numerous things that can be found in biomedical literature is taken on by BioNER. These entities include genes, diseases, chemical names, and medicinal names. The ever-increasing number of biological entities, their plethora of synonyms, the popularity of abbreviations, the use of lengthy entity descriptions, and the combination of letters, symbols, and punctuation within these texts all contribute to this complexity. BioNER, integral to numerous applications in the field of natural language processing, plays a pivotal role in biomedical literature mining [25, 67, 68]. One prominent application involves biomedical relation extraction, which seeks to uncover intricate relationships between diverse biomedical entities, such as diseases, genes, species, and chemicals. The accuracy and quality of downstream relation extraction tasks are directly contingent on the performance of BioNER systems [69]. Furthermore, BioNER significantly contributes to the identification and classification of drugs and their interactions, critical for applications like drug discovery, drug safety monitoring, and personalized medicine [70]. It also facilitates knowledge base completion, enriching biomedical knowledge bases by automatically extracting and categorizing pertinent entities from text. This enriched knowledge is then harnessed for a multitude of purposes, including semantic

search, question answering, and data integration. BioNER's utility extends to biomedical question answering, assisting in understanding and responding to complex biomedical queries by identifying and categorizing relevant entities from both the question and the relevant text [71]. In the realm of biomedical text summarization, BioNER proves invaluable in identifying and extracting key biomedical entities from texts, thereby enabling the generation of concise, informative summaries [71]. Moreover, it enhances the accuracy and relevance of search results in biomedical information retrieval systems, where it identifies and classifies pertinent entities in both the query and the document collection [51, 67, 71, 72]. A simplified representation for identifying a Named entity can be represented as shown in Algorithm 1

***Algorithm 1**: Named Entity Recognition and Classification using BERT Model.*

Input: Tokens: A sequence of tokens representing the input text

Output: Named Entity Labels: A list of predicted entity labels for each token in the input text

```
function Named_Entity_Recognition_And_Classification(tokens):
  Initialize Named_Entity_Labels as an empty list
  for each token t in a sequence of tokens do
    Calculate the probability distribution over possible entity labels:
      P(label | token, context) = fmodel(token, context)
    Apply softmax function to P(label | token, context)
    Assign the label with the highest probability for token t
    Append the predicted entity label to Named_Entity_Labels
  end for
  return Named_Entity_Labels
end function
```

The actual calculation of these probabilities can involve complex machine learning models such as CRFs, BiLSTM-CRFs, or Transformer-based models, which take into account the token, its context, and various features for labelling.

In Algorithm 1:

- P(label | token, context) represents the probability of a specific label given the token and its context.
- fmodel() represents the function that the machine learning model (*e.g.*, CRF, BiLSTM, Transformer) uses to predict this probability.
- The model represents the machine learning model that has been trained to predict entity labels.

SciSpaCy, a Python library and models tailored for practical biomedical and scientific text processing, makes extensive use of spaCy [73]. spaCy itself employs deep learning techniques and, with the release of version 3, has transitioned to using the Transformer architecture as its deep learning model. It offers a customizable architecture that accommodates various deep learning techniques such as LSTMs, CRFs, and transformers to suit specific tasks [74].

BioBERT, on the other hand, is a pre-trained biomedical language representation model based on the BERT architecture [75]. It has been fine-tuned on an extensive corpus of biomedical literature, incorporating articles from PubMed and PMC as shown in Fig. (**1**). The adaptability of this model enables it to perform very well in many biomedical activities, such as biomedical named entity identification, connection extraction, question answering, and drug discovery. Both BioBERT and SciSpaCy are open-source tools, rendering them important resources for NER.

In this work, we present significant contributions to the domain of Biomedical Named Entity Recognition (BioNER), with prior focus on the utilization and examination of two prominent models, SciSpaCy and BioBERT, addressing the intricate challenges inherent in BioNER.

1. Through meticulous training on diverse biological datasets, we systematically evaluate the performance of SciSpaCy and BioBERT across various domains. Our comprehensive assessment employs key metrics such as F1 scores and processing speed, providing a detailed understanding of their efficacy in BioNER tasks.

2. The study offers valuable insights for the selection of tools tailored to specific BioNER requirements. By optimizing Named Entity Recognition in the field of biomedical research, our work aims to contribute to the broader advancement of this critical domain

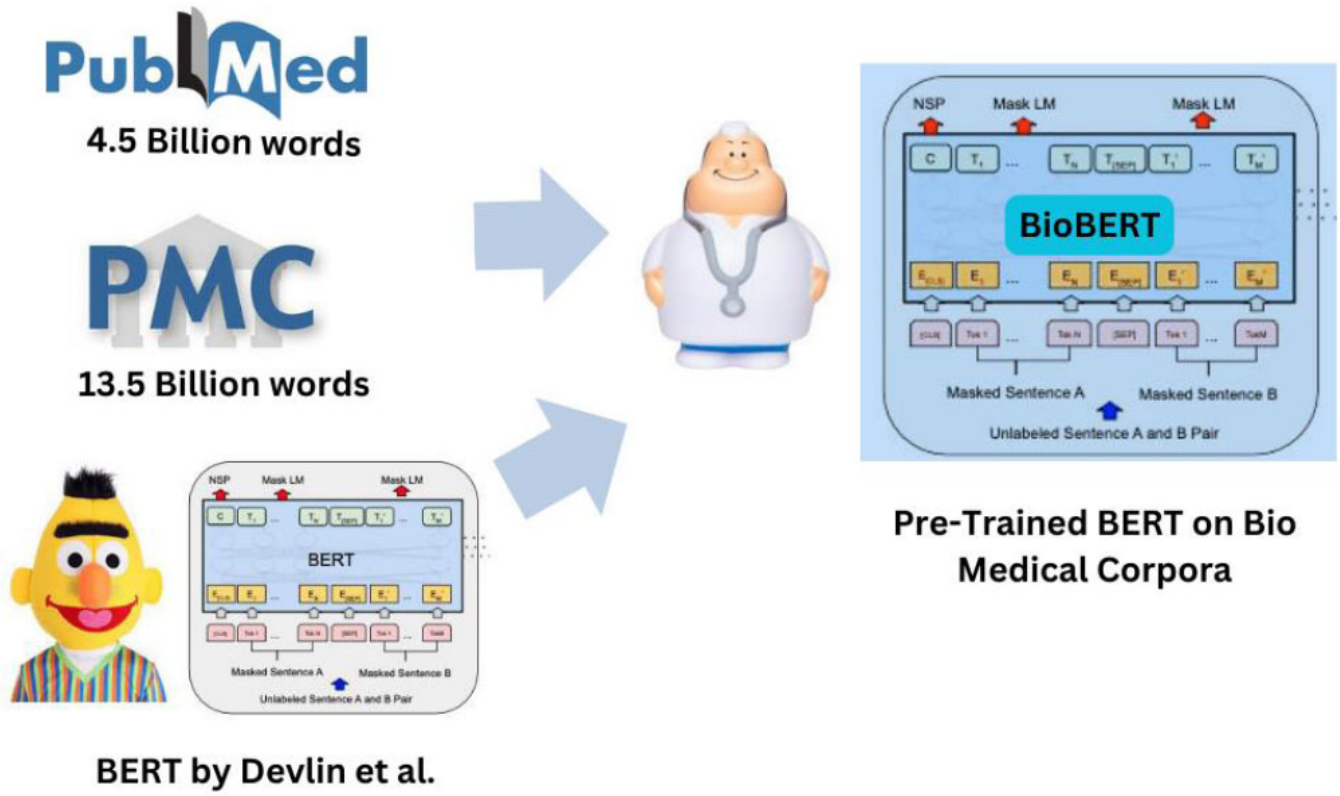


Fig. (1). BioBERT pretraining.

We have performed some literature surveys of the papers that collectively discuss various approaches to biomedical NER. Amith 2017 proposes an ontology-driven method that utilizes information extraction and features of ontologies to identify biomedical software names [76]. Al-Hegami 2017 focuses on the use of machine learning classifiers and a rich feature set to recognize biomedical entities, with the K-Nearest Neighbour classifier showing promising results [77]. Zhang 2013 presents an unsupervised approach to NER, using a noun phrase chunker and distributional semantics for entity extraction and classification [20].

Li 2023 proposes a fusion multi-features embedding method, combining deep contextual word-level features, local char-level features, and part-of-speech features to improve biomedical entity recognition [78]. Kaewphan *et al.*, 2018, published in the "Database Journal of Biological Databases Curation," shows a system for automatically identifying a wide range of biomedical entities within literature, boasting state-of-the-art performance. The study achieved the highest results in named entity recognition, underscoring its capability in biomedical entity recognition and normalization [79].

Priyanka *et al.*, 2021 provided a detailed survey of clinical Named Entity Recognition and Relationship Extraction (RE) techniques, addressing existing NLP models, performance, challenges, and future directions in the context of information extraction from clinical text, offering insights on current research and evaluation metrics [18].

Gong *et al.*, 2009 published in the "International Conference on BioMedical Engineering and Informatics," this study introduces a hybrid approach aimed at recognizing untagged biomedical entities. The approach demonstrates its utility in identifying biomedical entities in the GENIA 3.02 corpus, offering support for biologists in tagging biomedical entities within biomedical literature [80].

Kanimozhi & Manjula in 2018 performed a systematic review that focuses on the identification and classification of named entities within unstructured text documents. The goal is to extract valuable information from such texts by identifying and categorizing the named entities, a crucial step in extracting useful knowledge from unstructured textual data [81].

Overall, these papers highlight different techniques and algorithms for biomedical NER, showcasing the importance of effective natural language processing tools in organizing and extracting information from biomedical literature. Table **1** depicts the core methodology of the above-mentioned research works.

Table 1. NER tabular analysis.

S. No	Title	Authors	Year	Methodology
1.	"Knowledge-Based Approach for Named Entity Recognition in Biomedical Literature: A Use Case in Biomedical Software Identification"	Muhammad Amith, *et al*. [76]	2017	Ontology-driven method to identify familiar and unfamiliar software names.
2.	"A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set"	A. S. Al-Hegami, *et al*. [77]	2017	K-nearest neighbor trained with suitable features for recognizing biomedical named entities.
3.	"Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts"	Shaodian Zhang, *et al*. [20]	2013	Noun phrase chunker followed by a filter based on inverse document frequency for candidate entity extraction.
4.	"Biomedical named entity recognition based on fusion multi-features embedding"	Meijing Li, *et al*. [78]	2023	A proposed multi-feature embedding method with a positive effect on prediction results.
5.	"Wide-scope biomedical named entity recognition and normalization with CRFs, fuzzy matching and character level modelling"	S. Kaewphan, *et al*. [79]	2018	System for automatically identifying a multitude of biomedical entities from the literature with state-of-the-art performance.
6.	"A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts"	Priyanka *et al*. [18],	2021	Surveyed clinical Named Entity Recognition and Relationship Extraction (RE) techniques and addressed existing NLP models, performance, challenges.
7.	"A Hybrid Approach for Biomedical Entity Name Recognition"	Lejun Gong, *et al*. [80]	2009	A hybrid approach for recognizing untagged biomedical entities in GENIA 3.02 corpus.
8.	"A Systematic Review on Biomedical Named Entity Recognition"	U. Kanimozhi, *et al*. [81]	2018	Focus on identifying and classifying named entities for extracting useful information from unstructured text documents.

In addition to this study, there are several domains where Named Entity Recognition may be applied. Potential intersections and possibilities for collaboration or cross-pollination exist across the fields, especially in areas such as Visual Named Entity Recognition, cross-modal learning, semantic comprehension, and multi-modal information fusion.

1. Visual Named Entity Recognition (VNER): is a process that involves detecting and classifying named entities in text-based data. There is an increasing interest in Visual Named Entity Recognition, which refers to the process of recognising and categorising named items in photos or videos. The domain might possibly benefit from the application or adaptation of image processing techniques, particularly those that emphasise feature extraction and semantic interpretation.

2. Cross-Modal Learning: Cross-modal learning involves leveraging information from different modalities (*e.g.*, text and images) to improve performance on various tasks. Techniques developed for image processing, especially those involving attention mechanisms and feature extraction, could potentially be integrated with NER systems to improve entity recognition performance, especially in cases where textual context alone may not be sufficient.

3. Semantic Understanding: Both image processing and NER involve understanding the semantic content of data, albeit in different modalities. Techniques developed for image inpainting, super-resolution, or feature extraction may enhance the semantic understanding of textual data, which could indirectly benefit NER systems by providing richer contextual information.

4. Multi-Modal Information Fusion: Another potential linkage is through multi-modal information fusion techniques. Research in image processing often involves integrating information from different sources or modalities to improve the quality of image reconstruction or understanding. Similarly, integrating information from images (*e.g.*, scenes, objects) with textual data (*e.g.*, context, descriptions) could enhance NER systems' performance, especially in scenarios where textual and visual information are both available.

Some interesting works where NER can be combined are done by Yuantao Chen *et al*. in image processing techniques such as image inpainting and super-resolution. In a series of innovative contributions by Yuantao Chen, Runlong Xia, Kai Yang, and Ke Zou, their research endeavors have significantly advanced the field of image processing and restoration. In June 2023, they introduced "DARGS Image inpainting algorithm *via* deep attention residuals group and semantics," proposing a method that combines Semantic Priors Network, Deep Attention Residual Group, and Full-scale Skip Connection to effectively restore missing regions in images. Building on this foundation [82], in October 2023, they presented "GCAM: lightweight image inpainting *via* group convolution and attention mechanism," offering a lightweight approach that utilizes group convolution and rotating attention mechanism to enhance image inpainting while optimizing resource usage [83]. In January 2024, they unveiled "MFMAM: Image inpainting *via* multi-scale feature module with attention module," introducing a network utilizing multi-scale feature module and improved attention mechanisms for enhanced image inpainting, particularly focusing on texture and semantic detail preservation [84]. Their work continued with "DNNAM: Image inpainting algorithm *via* deep neural networks and attention mechanism," published in March 2024, where they introduced a method incorporating partial multi-scale channel attention mechanism and deep neural networks to improve image inpainting accuracy and quality [85]. Lastly, in July 2024, their paper titled "MICU: Image super-resolution *via* multi-level information compensation

and U-net" presented a novel approach for image super-resolution using multi-level information compensation and U-net architecture, significantly improving image reconstruction quality compared to existing methods [86]. Collectively, these papers underscore the author group's commitment to advancing image processing techniques with innovative methodologies and improved performance metrics. These works are very close in the possibilities for collaborating NER in areas such as Visual Named Entity Recognition, cross-modal learning, semantic comprehension, and multi-modal information fusion.

2. METHODOLOGY

The work aimed to rigorously evaluate the performance of BioBERT and SCISpaCy in biomedical Named Entity Recognition, considering the data's origin, preprocessing, training, and evaluation processes. The choice of these models and the training process was guided by the aim of achieving accurate and generalizable recognition of biomedical entities. These considerations are explained as follows:

2.1. Data Collection

The dataset utilized in this study was curated from diverse sources of biomedical Named Entity Recognition datasets. These datasets, including JNLPBA [15], BC4_CHEMD [87], BIONLP13CG [88], EXPERT 206 [70], and BC5CDR [89], were selected to encompass a wide range of biomedical entity types and domains, offering a comprehensive evaluation of BioNER models as shown in Table 2.

Entity Types and Datasets: The evaluation encompassed multiple datasets, each representing distinct biomedical entity types. The datasets included.

- JNLPBA: Targeted entities included DNA, CELL_TYPE, CELL_LINE, RNA, and PROTEIN.
- BC4_CHEMD: Focused on the CHEMICAL entity type.
- BIONLP13CG: Encompassed a diverse set of entity types, such as AMINO_ACID, ANATOMICAL_SYSTEM, CANCER, and more.
- EXPERT 206: Included GENE, DISEASE, and VARIANT as recognized entities.
- BC5CDR: Considered both CHEMICAL and DISEASE entities.

2.2. Data Preprocessing

2.2.1. Data Conversion for BioBERT

To prepare the data for training BioBERT, the datasets were converted into a BIO (Beginning-Inside-Outside) format. This involved annotating each token in the dataset with one of three labels: "B-" for the beginning of an entity, "I-" for internal parts of an entity, and "O" for tokens outside of any entity. This format enables BioBERT to understand and recognize the boundaries of biomedical entities within the text.

2.2.2. Data Conversion for SCISpaCy

For training SCISpaCy, the datasets were transformed into a format suitable for spaCy's NER model. The conversion process required structuring the data as a list of sentences, each followed by a list of (start, end, label) triples, where the triples define the start and end positions of the entity within the sentence and its associated label. For example, ("Tokyo Tower is 333m tall.", [(0, 11, "BUILDING")]) represents a sentence, the start and end positions of the entity "BUILDING," and its label.

2.3. Data Splitting

To facilitate model evaluation, the dataset was divided into training and testing subsets. The data was partitioned using an 80-20 split, allocating 80% of the data for training and 20% for testing. This ensures that the models are trained on a substantial portion of the data while preserving an independent dataset for assessing their generalization performance.

2.4. Hardware and Software Environment

Training of the model was carried out on carefully managed hardware consisting of Google Collab T4 GPUs. The environment had the essential programming languages, libraries, and frameworks for the construction and execution of the models, as well as particular GPU specifications for rapid training.

Table 2. Dataset analysis.

Datasets	Entity Type	No. of Entities
BIONLP13CG	Tissues	587
BIONLP13CG	Organisms	2,093
BIONLP13CG	Gene/Protein	7,908
EXPERT 206	Disease	7,390
EXPERT 206	Gene/Protein	7,393
BC5CDR	Disease	12,204
BC5CDR	Chemical	12,204
BC4_CHEMD	Chemical	84,249
JNLPBA	Cell Line	4,315

Table 3. NER results of SciSpacy and BioBERT.

Model	Entity Types	SCISpaCy F1 Score	BioBERT F1 Score
JNLPBA	"DNA, CELL_TYPE, CELL_LINE, RNA, PROTEIN"	73.1	73.67
BC4_CHEMD	"CHEMICAL"	85.1	86.07
BIONLP13CG	"AMINO_ACID, ANATOMICAL_SYSTEM, CANCER, CELL, CELLULAR_COMPONENT, DEVELOPING_ANATOMICAL_STRUCTURE, GENE_OR_GENE_PRODUCT, IMMATERIAL_ANATOMICAL_ENTITY, MULTI-TISSUE_STRUCTURE, ORGAN, ORGANISM, ORGANISM_SUBDIVISION, ORGANISM_SUBSTANCE, PATHOLOGICAL_FORMATION, SIMPLE_CHEMICAL, TISSUE"	78.13	86.07
EXPERT 206	"GENE, DISEASE, VARIANT"	91.08	91.23
BC5CDR	"CHEMICAL, DISEASE"	85.53	87.83

3. RESULTS AND DISCUSSION

3.1. Model Performance

3.1.1. JNLPBA Dataset

For the JNLPBA dataset, SCISpaCy and BioBERT both displayed competitive performance, with F1 scores of 73.1 and 73.67, respectively. These ratings were uniform across all entity types in this sample.

In the BC4 CHEMD dataset, which concentrates on CHEMICAL items, both models demonstrated excellent performance. BioBERT surpassed SCISpaCy with an F1 score of 86.07, while SCISpaCy received an F1 score of 85.1.

3.1.2. BIONLP13CG Dataset

The BIONLP13CG dataset featured a large variety of entity types, posing a challenge to the models' ability to recognise a wide range of biomedical items. BioBERT's F1 score of 86.07 was superior to SCISpaCy's score of 78.13.

The EXPERT 206 dataset, which consists of GENE, DISEASE, and VARIANT entities, produced high F1 scores for both models. SCISpaCy obtained an F1 value of 91.08, but BioBERT fared somewhat better with an F1 score of 91.23.

3.1.3. BC5CDR Dataset

For the BC5CDR dataset, which involves CHEMICAL and DISEASE entity recognition, both models displayed strong performance. SCISpaCy achieved an F1 score of 85.53, while BioBERT exhibited superior performance with an F1 score of 87.83. These results are showcased in Table **3**.

3.2. Inference and Performance Trade-offs

The evaluation of inference times revealed that SCISpaCy processed an average sentence in approximately 90 milliseconds, while BioBERT required an average of 278 milliseconds per sentence as shown in Fig. (**2**). It is important to note that the relatively longer inference time of BioBERT can be attributed to various internal tasks, including tokenizer loading and token classification.

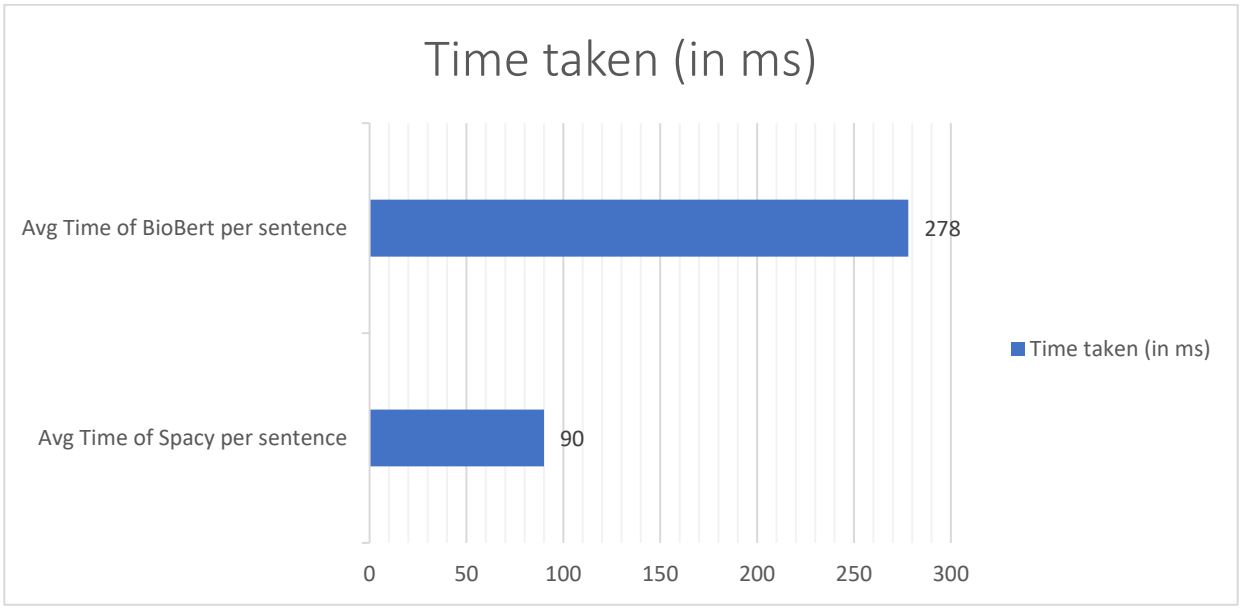


Fig. (2). Time complexity analysis of NER models.

However, BioBERT compensates for this by efficiently processing multiple sentences in parallel. This capability enhances throughput, making BioBERT a favorable choice in scenarios where parallelism is crucial.

4. DISCUSSION

Based on the results showcased in Table **3**. Some inferences have been deduced. These inferences can help to guide the selection of the appropriate model for specific biomedical NER tasks based on the dataset's entity types and performance requirements.

4.1. Entity Types and Specialization

The F1 scores vary across different entity types and datasets. For instance, in the JNLPBA dataset, DNA, CELL_TYPE, CELL_LINE, RNA, and PROTEIN have relatively similar F1 scores, indicating that the model's performance is consistent across these biomedical entity types. On the other hand, in the BIONLP13CG dataset, there is a wide range of entity types with varying F1 scores, suggesting that the model's performance is more specialized for some entity types.

4.2. BioBERT's Consistency

BioBERT consistently achieves higher F1 scores compared to SCISpaCy across all datasets and entity types. This indicates BioBERT's robustness and effectiveness in biomedical named entity recognition tasks.

4.3. High Performance for EXPERT 206

The EXPERT 206 dataset demonstrates the highest F1 scores for both models. This suggests that the entities in this dataset, such as Genes, Diseases, and variants, are well-recognized by both SCISpaCy and BioBERT.

4.4. Strong BioBERT Performance in BC5CDR

In the BC5CDR dataset, which focuses on Chemicals, BioBERT significantly outperforms SCISpaCy with an F1 score of 87.83, indicating its strength in recognizing chemical entities.

4.5. Balanced Performance in BC4_CHEMD

The BC4_CHEMD dataset, consisting of DISEASE and CHEMICAL entity types, shows similar and high F1 scores for both models. This suggests that both SCISpaCy and BioBERT are proficient in recognizing these types of entities.

4.6. Biomedical Specialization

The choice of model may depend on the specific biomedical entity types in the dataset. While BioBERT generally outperforms SCISpaCy, practitioners should consider the nature of the entities they are dealing with when selecting the most suitable model for their task.

4.7. Overall Robustness

Both SCISpaCy and BioBERT demonstrate strong performance in biomedical named entity recognition. However, BioBERT's consistently higher F1 scores and adaptability to various entity types make it a valuable choice for a wide range of biomedical applications.

4.8. Time Complexity on Same Hardware

On the same hardware configuration with a T4 GPU (utilizing Google Colab's default GPU settings), we observed that spaCy demonstrated a faster average inference time per sentence, approximately 50 milliseconds, compared to BioBERT's average speed of 178 milliseconds per sentence. It is worth noting that the apparent time difference in favor of spaCy may be attributed to its streamlined processing for individual sentences. However, it is important to consider that BioBERT exhibits a significant advantage when processing multiple sentences in parallel. This capability compensates for the additional time incurred by BioBERT in various internal tasks such as tokenizer loading and token classification. Consequently, BioBERT's suitability for parallel sentence processing makes it a strong contender in scenarios where high throughput and parallelism are critical.

CONCLUSION

SCISpaCy and BioBERT, along with other BioNER models, have demonstrated their capabilities, benefits, and relevant considerations. This study evaluated diverse datasets, and performance metrics, offering valuable insights to scholars and practitioners in the field of biological natural language processing. The study produced intriguing findings. The selection of a model such as SCISpaCy or BioBERT significantly impacts the detection of many biological events. BioBERT consistently achieved higher F1 scores than SCISpaCy across several datasets, entity types, and domains. The durability and adaptability of BioBERT in the biomedical field make it a highly promising choice for BioNER.The datasets exhibited notable variations in F1 scores across different item categories. The F1 ratings for both models in the EXPERT 206 dataset of Genes, Diseases, and Variants indicate that SCISpaCy and BioBERT effectively recognise and categorise these entity types. The BIONLP13CG dataset exhibited a diverse range of entity kinds, each with distinct F1 scores. The performance of the model may be tailored to specific entities. Parallel processing is an additional advantage of BioBERT. Although BioBERT may have longer phrase inference times, it has the ability to process several sentences simultaneously, which makes it an excellent choice for high-throughput applications. It is essential to assess the different categories within the dataset and determine the optimal trade-off between efficiency and precision when recognising biological entities. The selection of the model should be guided by the intrinsic aspects of the work. Our analysis demonstrates the significance of BioNER in the field of biomedical research and the capabilities of current models. Before making a decision, it is important to thoroughly comprehend the specific criteria and objectives of each biological Named Entity Recognition task, and then select either SCISpaCy, BioBERT, or a combination of both accordingly. The advancement of NLP models in the field of biomedicine provides professionals with a diverse

array of tools to effectively extract pertinent information from texts related to biology.

LIMITATIONS, FUTURE WORK AND RESEARCH DIRECTIONS

To improve the thoroughness of our assessments, it is crucial to further investigate domain-specific algorithms for Named Entity Recognition in the context of future studies. In order to achieve this objective, we suggest a deliberate enlargement of our study scope to encompass notable algorithms such as Bert-PKD, CollaboNet, TinyBert, BERN2, and other state-of-the-art techniques. Integrating these algorithms into our future endeavours will enhance the strength and comprehensiveness of our evaluation system. The limitations in the field of Biomedical Named entity recognition include data availability, expert annotation requirements, and the vast space of biomedical concepts. Our objective is to analyse a wider range of NER algorithms in a systematic manner, in order to reveal subtle variations in performance and find methods that may have distinct benefits in particular biological scenarios. This strategic growth is in line with our dedication to developing the area of Biomedical Named Entity Recognition and guaranteeing that our research offers practical insights for researchers, practitioners, and professionals in the biomedical sector. Integrating new algorithms will enhance the comprehensiveness of our studies and promote a more nuanced comprehension of the advantages and constraints of different techniques. To summarise, more study on domain-specific algorithms for NER is crucial. There are potential intersections and opportunities for collaboration or cross-pollination between the fields, particularly in areas such as Visual Named Entity Recognition, cross-modal learning, semantic understanding, and multi-modal information fusion. We are committed to expanding the reach and significance of our findings in future studies.

LIST OF ABBREVIATIONS

BioNER	=	Biological Named Entity Recognition
NER	=	Named Entity Recognition
IE	=	Information Extraction
NLP	=	Natural Language Processing
CRF	=	Conditional Random Fields

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIAL

All the data and supporting information is provided within the article.

FUNDING

None.

CONFLICT OF INTEREST

The authors declared no conflict of interest financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, and Z. Lu, "BioRED: A rich biomedical relation extraction dataset", *Brief. Bioinform.,* vol. 2022, no. 5, pp. 1-12, 2022.

[2] Q. Xi, Y. Ren, S. Yao, G. Wu, G. Miao, and Z. Zhang, "Chinese named entity recognition", *Applicat. Challeng.,* vol. 12647, pp. 51-81, 2021. [http://dx.doi.org/10.1007/978-3-030-71590-8_4]

[3] A. Ahmed, A. Abbasi, and C. Eickhoff, "Benchmarking modern named entity recognition techniques for free-text health record deidentification", *AMIA Jt Summits Transl Sci Proc,* vol. 2021, pp. 102-111, 2021.

[4] R. Hema, and A. Devi, "Chemical named entity recognition using deep learning techniques", *Deep Natural Language Processing and AI Applications for Industry* pp.59-73, 2021. [http://dx.doi.org/10.4018/978-1-7998-7728-8.ch004]

[5] K. Dawar, A.J. Samuel, and R. Alvarado, "Comparing topic modeling and named entity recognition techniques for the semantic indexing of a landscape architecture textbook", *2019 Syst. Inf. Eng. Des. Symp.,* 2019pp. 1-6

[6] S. Tedeschi, and R. Navigli, "MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and Disambiguation)", *Findings of the Association for Computational Linguistics: NAACL* Seattle, United States pp.801-812, 2022.

[7] A. Groza, "Detecting fake news for the new coronavirus by reasoning on the Covid-19 ontology", *arXiv:2004.12330,* 2020.

[8] H. Liu, Z. Sun, and F. Ning, "Named entity recognition method for cnc machine tool design knowledge text", *2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)* 11-12 December 2022, Dalian, China, pp.1452-1455, 2022.

[9] S.R. Kundeti, J. Vijayananda, S. Mujjiga, and M. Kalyan, "Clinical named entity recognition: Challenges and opportunities", *2016 IEEE International Conference on Big Data (Big Data)* 05-08 December 2016, Washington, DC, USA, pp.1937-1945, 2016.

[10] M.C. Durango, T.E.A. Silva, and O.A. Duque, "Named entity recognition in electronic health records: A methodological review", *Healthc. Inform. Res.,* vol. 29, no. 4, pp. 286-300, 2023. [http://dx.doi.org/10.4258/HIR.2023.29.4.286] [PMID: 37964451]

[11] N. Perera, M. Dehmer, and F. Emmert-Streib, "Named entity recognition and relation detection for biomedical information extraction", *Front Cell Dev Biol,* vol. 8, p. 673, 2020.

[12] M. Ehrmann, A. Hamdi, E.L. Pontes, M. Romanello, and A. Doucet, "Named entity recognition and classification in historical documents: A survey", *ACM Comput. Surv.,* vol. 56, no. 2, 2023. [http://dx.doi.org/10.1145/3604931]

[13] H. Cho, and H. Lee, "Biomedical named entity recognition using deep neural networks with contextual information", *BMC Bioinformat.,* vol. 20, no. 1, p. 735, 2019. [http://dx.doi.org/10.1186/S12859-019-3321-4/FIGURES/1] [PMID: 31881938]

[14] V. Kocaman, and D. Talby, "Accurate clinical and biomedical named entity recognition at scale", *Softw. Impacts,* vol. 13, p. 100373, 2022. [http://dx.doi.org/10.1016/J.SIMPA.2022.100373]

[15] M-S. Huang, P-T. Lai, R.T-H. Tsai, and W-L. Hsu, "Revised JNLPBA corpus: A revised version of biomedical NER corpus for relation extraction task", *Brief. Bioinform.,* vol. 21, no. 6, pp. 2219-2238, 2020. [http://dx.doi.org/10.1093/bib/bbaa054] [PMID: 32602538]

[16] K. Wang, "NERO: A biomedical named-entity (recognition) ontology with a large, annotated corpus reveals meaningful associations through text embedding", *NPJ Syst Biol Appl,* vol. 7, p. 38, 2021.

[17] A. Goyal, V. Gupta, and M. Kumar, "Recent named entity recognition and classification techniques: A systematic review",

Comput. Sci. Rev., vol. 29, pp. 21-43, 2018.
[http://dx.doi.org/10.1016/J.COSREV.2018.06.001]

[18] P. Bose, S. Srinivasan, W.C. Sleeman, J. Palta, R. Kapoor, and P. Ghosh, "A survey on recent named entity recognition and relationship extraction techniques on clinical texts", *Appl. Sci.,* vol. 11, no. 18, p. 8319, 2021.
[http://dx.doi.org/10.3390/app11188319]

[19] V. Moscato, M. Postiglione, and G. Sperlí, "Few-shot named entity recognition: Definition, taxonomy and research directions", *ACM Trans. Intell. Syst. Technol.,* vol. 14, no. 5, pp. 1-46, 2023.
[http://dx.doi.org/10.1145/3609483]

[20] S. Zhang, and N. Elhadad, "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts", *J. Biomed. Inform.,* vol. 46, no. 6, pp. 1088-1098, 2013.
[PMID: 23954592]

[21] Y. Tian, W. Shen, Y. Song, F. Xia, M. He, and K. Li, "Improving biomedical named entity recognition with syntactic information", *BMC Bioinformat.,* vol. 21, no. 1, p. 539, 2020.
[http://dx.doi.org/10.1186/s12859-020-03834-6] [PMID: 33238875]

[22] J. Kocerka, M. Krześlak, and A. Gałuszka, "Ontology extraction from software requirements using named-entity recognition", *Adv. Sci. Technol. Res. J.,* vol. 16, no. 3, pp. 207-212, 2022.
[http://dx.doi.org/10.12913/22998624/149941]

[23] G.L. Ciampaglia, P. Shiralkar, L.M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks", *PLoS One,* vol. 10, no. 6, p. e0128193, 2015.
[http://dx.doi.org/10.1371/JOURNAL.PONE.0128193] [PMID: 26083336]

[24] R. Meyes, M. Lu, C.W. de Puiseau, and T. Meisen, "Ablation studies in artificial neural networks", *arXiv:1901.08644,* 2019.

[25] Y. Xiong, "Improving deep learning method for biomedical named entity recognition by using entity definition information", *BMC Bioinformat.,* vol. 22, no. S1, p. 600, 2021.

[26] A. Goyal, M. Kumar, and V. Gupta, "Named entity recognition: Applications, approaches and challenges", *IJARCCE,* vol. 6, no. 2, pp. 259-262, 2017.

[27] L. Chang, R. Zhang, J. Lv, W. Zhou, and Y. Bai, "A review of biomedical named entity recognition", *J. Comput. Methods Sci. Eng.,* vol. 22, no. 3, pp. 893-900, 2022.
[http://dx.doi.org/10.3233/JCM-225952]

[28] P. Kalamkar, A. Agarwal, A. Tiwari, S. Gupta, S. Karn, and V. Raghavan, "Named entity recognition in indian court judgments", *arXiv:2211.03442,* 2022.

[29] B. Yaman, M. Pasin, and M. Freudenberg, "Interlinking SciGraph and DBpedia datasets using link discovery and named entity recognition techniques", *Open Access Ser. Informat. (OASIcs),* vol. 70, pp. 15:1-15:8, 2019.

[30] L.S. Cocca, B. Maier, C. Nawroth, P. Kevitt, and M. Hemmje, "Named entity recognition for the extraction of emerging technological knowledge from medical literature", *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management,* 2022pp. 101-108

[31] M. Liu, Z. Tu, T. Zhang, T. Su, X. Xu, and Z. Wang, "LTP: A new active learning strategy for CRF-based named entity recognition", *Neural Process. Lett.,* vol. 54, no. 3, pp. 2433-2454, 2022.
[http://dx.doi.org/10.1007/s11063-021-10737-x]

[32] H. Luo, and B. Gao, "Development of smart wearable sensors for life healthcare", *Eng. Regen.,* vol. 2, pp. 163-170, 2021.
[http://dx.doi.org/10.1016/J.ENGREG.2021.10.001]

[33] Z. Li, S. Zhang, Y. Song, and J. Park, "Extrinsic factors affecting the accuracy of biomedical NER", *arXiv:2305.18152,* 2023.

[34] N. Garg, "Research Proposal Research Question " Can Named Entities improve", Available from: https://www.semanticscholar.org/paper/Research-Proposal-Research-Question-

[35] J. Barua, and D. Patel, "Named entity classification using search engine's query suggestions", *European Conference on Information Retrieval,* vol. 2017, 2017.

[36] C. Sabty, I. Omar, F. Wasfalla, M. Islam, and S. Abdennadher, "Data augmentation techniques on arabic data for named entity recognition", *Procedia Comput. Sci.,* vol. 189, pp. 292-299, 2021.
[http://dx.doi.org/10.1016/j.procs.2021.05.092]

[37] M.H. Khanam, M.A. Khudhus, and M.S.P. Babu, "Named entity recognition using machine learning techniques for telugu language", *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)* 26-28 August 2016, Beijing, China, 2016.

[38] W. Shishah, "Fake news detection using BERT model with joint learning", *Arab. J. Sci. Eng.,* vol. 46, pp. 9115-9127, 2021.
[http://dx.doi.org/10.1007/s13369-021-05780-8]

[39] J. Kalluru, "Enhancing data accuracy and efficiency: An overview of fuzzy matching techniques", *Int. J. Sci. Res.,* vol. 12, no. 8, pp. 685-690, 2023.

[40] E. Tokarchuk, D. Thulke, W. Wang, C. Dugast, and H. Ney, "Investigation on data adaptation techniques for neural named entity recognition", *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop* pp.1-15, 2021.

[41] G. S. Mahalakshmi, and A.A. L. Adline, "Boosting formal concept analysis based definition extraction via named entity recognition", In: *Smart Innovation, Systems and Technologies,* Springer: New Delhi, 2015.L. Jain, H. Behera, J. Mandal, and D. Mohapatra,

[42] T. Mehmood, I. Serina, A. Lavelli, L. Putelli, and A. Gerevini, "On the use of knowledge transfer techniques for biomedical named entity recognition", *Future Internet,* vol. 15, no. 2, p. 79, 2023.
[http://dx.doi.org/10.3390/fi15020079]

[43] L. Stepanyan, "Automated custom named entity recognition and disambiguation", Available from: https://api.semanticscholar.org/CorpusID:219618671

[44] M. Pajas, A. Radovan, and I.O. Biškupić, "Multilingual named entity recognition solution for optimizing parcel delivery in online commerce: Identifying person and organization names", *2023 46th MIPRO ICT and Electronics Convention (MIPRO)* 22-26 May 2023, Opatija, Croatia, pp.1119-1124, 2023.
[http://dx.doi.org/10.23919/MIPRO57284.2023.10159789]

[45] C. Jilek, M. Schröder, R. Novik, S. Schwarz, H. Maus, and A. Dengel, "Inflection-tolerant ontology-based named entity recognition for real-time applications", *arXiv:1812.02119,* 2019.

[46] X. Jiang, and T. Ouyang, "End-to-end speech to named entity recognition system", Available from: https://www.semanticscholar.org/paper/End-to-End-Speech-to-Named-Entity-Recognition-Jiang-Ouyang/b2c43c47151c339c92b8e1be629c152626afb90a

[47] Q. Wei, Z. Ji, Z. Li, J. Du, J. Wang, J. Xu, Y. Xiang, F. Tiryaki, S. Wu, Y. Zhang, C. Tao, and H. Xu, "A study of deep learning approaches for medication and adverse drug event extraction from clinical text", *J. Am. Med. Inform. Assoc.,* vol. 27, no. 1, pp. 13-21, 2020.
[http://dx.doi.org/10.1093/jamia/ocz063] [PMID: 31135882]

[48] Arakelyan, "Automated custom named entity recognition and disambiguation", Available from: https://www.semanticscholar.org/paper/Automated-Custom-Named-Entity-Recognition-and-Stepanyan/24d73ef1aa9f2fb7c1651be67b4f3e40b55ff31e

[49] B. Powley, and R. Dale, "High accuracy citation extraction and named entity recognition for a heterogeneous corpus of academic papers", Available from: https://www.semanticscholar.org/paper/High-accuracy-citation-extraction-and-named-entity-Powley-Dale/e478b6069a7fe2dae89673553edad449ec6c329b

[50] G. Alfattni, M. Belousov, N. Peek, and G. Nenadic, "Extracting drug names and associated attributes from discharge summaries: Text mining study", *JMIR Med. Inform.,* vol. 9, no. 5, p. e24678, 2021.
[http://dx.doi.org/10.2196/24678] [PMID: 33949962]

[51] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, "Biomedical named entity recognition using BERT in the machine

reading comprehension framework", *J. Biomed. Inform.,* vol. 118, p. 103799, 2021.
[http://dx.doi.org/10.1016/j.jbi.2021.103799] [PMID: 33965638]

[52] T. Zhang, "BDANN: BERT-based domain adaptation neural network for multi-modal fake news detection", *2020 International Joint Conference on Neural Networks (IJCNN)* 19-24 July 2020, Glasgow, UK, 2020.

[53] T.M. Luu, R. Phan, R. Davey, and G. Chetty, "A multilevel NER framework for automatic clinical name entity recognition", *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* 18-21 November 2017, New Orleans, LA, USA, pp.1134-1143, 2017.

[54] U. Qazi, M. Imran, and F. Ofli, "GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information", *arXiv:2005.11177,* 2020.

[55] S. Raza, D.J. Reji, F. Shajan, and S.R. Bashir, "Large-scale application of named entity recognition to biomedicine and epidemiology", *PLOS Digit. Heal.,* vol. 1, no. 12, p. e0000152, 2022.
[http://dx.doi.org/10.1371/journal.pdig.0000152] [PMID: 36812589]

[56] H. Gobbi, and M. De Brot, "Papillary tumors of the breast", In: S. Stolnicu, I. Alvarado-Cabrero, Eds., *Practical Atlas of Breast Pathology,* Springer: Cham, 2018.

[57] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news", *Proceedings of the 27th International Conference on Computational Linguistics* Santa Fe, New Mexico, USA, pp.3391–3401, 2018.

[58] M. Linzer, E.H. Yang, N.A. Estes III, P. Wang, V.R. Vorperian, and W.N. Kapoor, "Diagnosing syncope. Part 1: Value of history, physical examination, and electrocardiography", *Ann. Intern. Med.,* vol. 126, no. 12, pp. 989-996, 1997.
[http://dx.doi.org/10.7326/0003-4819-126-12-199706150-00012] [PMID: 9182479]

[59] N.S. Pagad, and N. Pradeep, "Clinical named entity recognition methods: An overview", In: *Advances in Intelligent Systems and Computing,* vol. 1388. Springer: Singapore, 2022, pp. 151-165.
[http://dx.doi.org/10.1007/978-981-16-2597-8_13]

[60] S. Raza, and B. Schwartz, "Detecting biomedical named entities in COVID-19 texts", Available from: https://www.semanticscholar.org/paper/Detecting-Biomedical-Named-Entities-in-COVID-19-Raza-Schwartz/d38a9d72a347c404f69097be57bec6967b2e5bb1

[61] S.I. Bedmar, C.D. Perdonas, and G.S. Aspizua, "Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts", *BMC Bioinformat.,* vol. 23, no. 1, p. 263, 2021.

[62] H-J. Dai, C-H. Wei, H-Y. Kao, R-L. Liu, R.T-H. Tsai, and Z. Lu, "Text mining for translational bioinformatics", *BioMed Res. Int.,* vol. 2015, p. 368264, 2015.
[http://dx.doi.org/10.1155/2015/368264] [PMID: 26380272]

[63] L-H. Lee, C-Y. Chen, L-C. Yu, and Y-H. Tseng, "Overview of the ROCLING 2022 shared task for chinese healthcare named entity recognition", Available from: https://www.semanticscholar.org/paper/Overview-of-the-ROCLING-2022-Shared-Task-for-Named-Lee-Chen/0f3b2ac0d7aa2c5d4c50b5bd066b94abbf53d0f1

[64] B. Tang, D. Jiang, Q. Chen, X. Wang, J. Yan, and Y. Shen, "De-identification of clinical text via Bi-LSTM-CRF with neural language models", *AMIA Annu Symp Proc.,* vol. 2019, pp. 857-863, 2019.

[65] C. Xarhoulacos, A. Anagnostopoulou, G. Stergiopoulos, and D. Gritzalis, "Misinformation vs. situational awareness : The art of deception and the need for cross-domain detection", *Sensors,* vol. 21, no. 16, p. 5496, 2021.

[66] A. Radford, "Better language models and their implications", Available from: https://openai.com/blog/better-language-models/

[67] H. Kim, and J. Kang, "How do your biomedical named entity recognition models generalize to novel entities?", *IEEE Access,* vol. 10, pp. 31513-31523, 2022.
[http://dx.doi.org/10.1109/ACCESS.2022.3157854] [PMID: 35582496]

[68] M. Alzantot, Y. Sharma, A. Elgohary, B-J. Ho, M.B. Srivastava, and K-W. Chang, "Generating natural language adversarial examples", Available from: https://github.com/nesl/nlp_

[69] Z. Zhang, and A.L.P. Chen, "Biomedical named entity recognition with the combined feature attention and fully-shared multi-task learning", *BMC Bioinformatics,* vol. 23, no. 1, p. 458, 2022.
[http://dx.doi.org/10.1186/s12859-022-04994-3] [PMID: 36329384]

[70] P-H. Li, T.F. Chen, J.Y. Yu, S.H. Shih, C.H. Su, Y.H. Lin, H.K. Tsai, H.F. Juan, C.Y. Chen, and J.H. Huang, "pubmedKB: An interactive web server for exploring biomedical entity relations in the biomedical literature", *Nucleic Acids Res.,* vol. 50, no. W1, pp. W616-W622, 2022.
[http://dx.doi.org/10.1093/nar/gkac310] [PMID: 35536289]

[71] S. Jansen, "Who's who and what's what: Advances in biomedical named entity recognition (BioNER)", Available from: https://towardsdatascience.com/whos-who-and-what-s-what-advances-in-biomedical-named-entity-recognition-bioner-c42a3f63334c

[72] L. Luo, C-H. Wei, P-T. Lai, R. Leaman, Q. Chen, and Z. Lu, "AIONER: All-in-one scheme-based biomedical named entity recognition using deep learning", *Bioinformatics,* vol. 39, no. 5, p. btad310, 2023.
[http://dx.doi.org/10.1093/bioinformatics/btad310] [PMID: 37171899]

[73] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and robust models for biomedical natural language processing", *arXiv:1902.07669,* 2019.

[74] "Industrial-strength natural language processing", Available from: https://spacy.io/

[75] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics,* vol. 36, no. 4, pp. 1234-1240, 2020.
[http://dx.doi.org/10.1093/bioinformatics/btz682] [PMID: 31501885]

[76] M. Amith, Y. Zhang, H. Xu, and C. Tao, "Knowledge-based approach for named entity recognition in biomedical literature: A use case in biomedical software identification", In: S. Benferhat, K. Tabia, M. Ali, Eds., *Advances in Artificial Intelligence: From Theory to Practice,* vol. Vol. 10351. Springer International Publishing: Cham, 2017, pp. 386-395.

[77] A.S. Al-Hegami, A.M.F. Othman, and F.T. Bagash, "A biomedical named entity recognition using machine learning classifiers and rich feature set", *Int. J. Comput. Sci. Netw. Secur.,* vol. 17, no. 1, p. 170, 2017.

[78] M. Li, H. Yang, and Y. Liu, "Biomedical named entity recognition based on fusion multi-features embedding", *Technol Health Care.,* vol. 31, no. S1, pp. 111-121, 2023.

[79] S. Kaewphan, K. Hakala, N. Miekka, T. Salakoski, and F. Ginter, "Wide-scope biomedical named entity recognition and normalization with CRFs, fuzzy matching and character level modeling", *Database,* vol. 2018, pp. 1-10, 2018.
[PMID: 30239666]

[80] L-J. Gong, Y. Yuan, Y-B. Wei, and X. Sun, "A hybrid approach for biomedical entity name recognition", *2009 2nd International Conference on Biomedical Engineering and Informatics* 17-19 October 2009, Tianjin, China, 2009.

[81] U. Kanimozhi, and D. Manjula, "A systematic review on biomedical named entity recognition", In: *Data Science Analytics and Applications,* Springer, 2018, pp. 19-37.

[82] Y. Chen, R. Xia, K. Yang, and K. Zou, "DARGS: Image inpainting algorithm via deep attention residuals group and semantics", *J. King Saud Univ. Comput. Inf. Sci.,* vol. 35, no. 6, p. 101567, 2023.
[http://dx.doi.org/10.1016/J.JKSUCI.2023.101567]

[83] Y. Chen, R. Xia, K. Yang, and K. Zou, "GCAM: Lightweight image inpainting via group convolution and attention mechanism", *Int. J. Mach. Learn. Cybern.,* no. Oct, pp. 1-11, 2023.
[http://dx.doi.org/10.1007/S13042-023-01999-Z/METRICS] [PMID:

37360881]
[84] Y. Chen, R. Xia, K. Yang, and K. Zou, "MFMAM: Image inpainting via multi-scale feature module with attention module", *Comput. Vis. Image Underst.,* vol. 238, p. 103883, 2024. [http://dx.doi.org/10.1016/J.CVIU.2023.103883]
[85] Y. Chen, R. Xia, K. Yang, and K. Zou, "DNNAM: Image inpainting algorithm via deep neural networks and attention mechanism", *Appl. Soft Comput.,* vol. 154, p. 111392, 2024. [http://dx.doi.org/10.1016/J.ASOC.2024.111392]
[86] Y. Chen, R. Xia, K. Yang, and K. Zou, "MICU: Image super-resolution via multi-level information compensation and U-net", *Expert Syst. Appl.,* vol. 245, p. 123111, 2024. [http://dx.doi.org/10.1016/J.ESWA.2023.123111]
[87] "Named Entity Recognition (NER) on BC4CHEMD", Available from: https://paperswithcode.com/sota/named-entity- recognition-on-bc4chemd
[88] AI Datasets. Available from: https://www.ncbi.nlm.nih.gov/research/bionlp/Data/
[89] "BC5CDR (BioCreative V CDR corpus)", Available from: https://paperswithcode.com/dataset/bc5cdr

Analytical and Bioanalytical Chemistry
https://doi.org/10.1007/s00216-024-05379-6

RESEARCH PAPER

Fabrication of rGO-decorated hBNNS hybrid nanocomposite via organic–inorganic interfacial chemistry for enhanced electrocatalytic detection of carcinoembryonic antigen

Kanika Sharma[1] · Nitin K. Puri[1] · Bharti Singh[1]

Received: 22 March 2024 / Revised: 12 May 2024 / Accepted: 15 May 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2024

Abstract
Organic–inorganic hybrid nanocomposites (OIHN), with tailored surface chemistry, offer ultra-sensitive architecture capable of detecting ultra-low concentrations of target analytes with precision. In the present work, a novel nano-biosensor was fabricated, acquainting dynamic synergy of reduced graphene oxide (rGO) decorated hexagonal boron nitride nanosheets (hBNNS) for detection of carcinoembryonic antigen (CEA). Extensive spectroscopic and microscopic analyses confirmed the successful hydrothermal synthesis of cross-linked rGO-hBNNS nanocomposite. Uniform micro-electrodes of rGO-hBNNS onto pre-hydrolyzed ITO were obtained via electrophoretic deposition (EPD) technique at low DC potential (15 V). Optimization of antibody incubation time, pH of supporting electrolyte, and immunoelectrode preparation was thoroughly investigated to enhance nano-biosensing efficacy. rGO-modified hBNNS demonstrated 29% boost in electrochemical performance over bare hBNNS, signifying remarkable electro-catalytic activity of nano-biosensor. The presence of multifunctional groups on the interface facilitated stable crosslinking chemistry, increased immobilization density, and enabled site-specific anchoring of Anti-CEA, resulting in improved binding affinity. The nano-biosensor demonstrated a remarkably low limit of detection of 5.47 pg/mL ($R^2 = 0.99963$), indicating exceptional sensitivity and accuracy in detecting CEA concentrations from 0 to 50 ng/mL. The clinical evaluation confirmed its exceptional shelf life, minimal cross-reactivity, and robust recovery rates in human serum samples, thereby unraveling the potential for early, highly sensitive, and reliable CEA detection.

Keywords Organic–inorganic hybrid nanocomposite · Reduced graphene oxide (rGO) · Hexagonal boron nitride nanosheets (hBNNS) · Non-small cell lung cancer (NSCLC) · Carcinoembryonic antigen (CEA)

Introduction

In today's global healthcare landscape, lung cancer, particularly, non-small cell lung cancer (NSCLC), remains a formidable adversary, ranking as one of the leading causes of premature mortality worldwide [1, 2]. With NSCLC accounting for a substantial portion of lung cancer cases (~ 80–85%), urgent attention is imperative to confront the escalating mortality rates associated with this disease. In response to this upsurge, countries across diverse levels of human development are intensifying efforts to bolster prevention rates [3]. Therefore, priorities are being set on developing advanced nano-biosensors that are feasible and can comprehensibly detect proliferating malignant cells at the nascent stage, thereby mitigating the alarming toll of premature mortality. Research consistently emphasizes the significance of over-expressed oncofetal biomarkers in cancer initiation and progression [4, 5]. Notably, monitoring elevated levels of CEA (molecular weight, 200 kDa), typically above 5 ng/mL in blood serum, emerges as a critical tool for early detection of NSCLC [6]. An elevated concentration of protein biomarkers in serum is considered an effective way of differentiating a diseased person from a normal person, estimating quantitative progression in the tumor, and monitoring the recovery rate after treatment.

Published in the topical collection *Emerging Trends in Electrochemical Analysis* with guest editors Sabine Szunerits, Wei Wang, and Adam T. Woolley.

✉ Nitin K. Puri
nitinkumarpuri@dtu.ac.in; nitinpuri2002@yahoo.co.in

[1] Nanomaterials Research Laboratory (NRL), Department of Applied Physics, Delhi Technological University (DTU), Shahbad Daulatpur, Bawana Road, Delhi 110042, India

Published online: 15 June 2024

Traditional methods for CEA detection, like immunohistochemistry, radioimmunoassay, polymerase chain reaction, and enzyme-linked immunoassay, present numerous challenges that include susceptibility to interference, time-intensive procedures, and potential hazards, demand intricate experimental setup, necessitate meticulous waste management, and often yield inefficient outcomes [7]. Realizing the inevitable exigency and limitations of existing diagnostic tools, electrochemical nano-biosensors are being developed on a large scale that can provide highly sensitive response, simple instrumentation, faster monitoring time, miniaturized device, inexpensive charges, and good portability [7, 8]. Electrochemical immunosensors exhibit a non-enzymatic nature that vanishes intervention of any additional modification in the form of the label, mediator, or substrate in buffer solution. This bypasses the lower binding efficiency caused by steric hindrance at the interface and non-specific binding of antibody-antigen complexes [7].

Thus, this work focuses on directly conjugating fragment crystallizable (Fc) receptors to nano-biosensor interface with strong adsorption affinity to achieve enhanced electro-catalytic efficiency, improved diagnostic value, and high chemical stability together with cost-effective fabrication. Nano-engineering innovations have facilitated the fabrication of bio-analytical devices through defect induction, functionalization, structural and morphological architecture modulation, and hybrid composite formation [9]. Despite their promising potential, structural, morphological, and defect engineering methods often encounter challenges in ambient environments such as instability, degradation, complex fabrication processes, difficulty in maintaining precise control over material properties and compromised chemical as well as biological compatibility of nanomaterials, which can hinder their practical applications. Therefore, Synergistically coupled organic–inorganic hybrid nanocomposites (OIHN), enhanced by functionalization, effectively overcome challenges associated with the aforementioned techniques by harnessing unique characteristics enabled by host–guest chemistry between biphasic nano-confined parent materials [8, 10]. Organic materials are typically lightweight and fragile, with a lack of chemical and thermal stability at elevated temperatures. Conversely, inorganic counterparts, though robust, often lack diverse functionalities, demand intricate synthesis, and display limited catalytic efficiency. OIHN formation ingeniously tackles challenges from both counterparts by meticulously engineering interfacial architecture. This approach not only enhanced mechanical, thermal, and chemical stability but also tailors electrical and optical properties. Moreover, it ensures biocompatibility, facilitates surface functionalization, and unlocks synergistic catalytic activity.

Two-dimensional (2D) nanomaterials emerge as a compelling candidate, offering abundant surface-active sites for biomolecule loading, enhanced catalytic efficiency, extraordinary bond tailoring capabilities, and enabled bandgap tunability [11]. Layered materials serve as scalable sources for achieving 2D morphology in OIHN. Layered nanostructures, abundant in both organic and inorganic material families, possess unique properties absent in their counterparts due to their molecular pairing. Among layered organic nanomaterials, reduced graphene oxide (rGO) stands out for its high aspect ratio, providing extensive surface-active adsorption sites, semiconducting bandgap enhancing electrocatalytic activity, and high defect density facilitating enhanced bio-catalytic activity [12, 13]. Complementing the performance of organic counterparts is hexagonal boron nitride (hBN), also known as inorganic graphite. With lattice parameters comparable to rGO, hBN minimizes strain-induced deformation, offering remarkable mechanical strength and thermal and chemical stability, making it an ideal candidate for hybrid nanocomposite formation [14].

Within the domain of organic nanomaterials, reduced graphene oxide (rGO) exhibits a distinctive electronic structure featuring uniformly distributed electrons across each carbon atom, underscoring the presence of numerous redox sites. These abundant redox centers ensure enhanced electron transfer kinetics, making rGO a valuable candidate for OIHN-based nano-biosensors, while electrons in hexagonal boron nitride (hBN) gravitate towards comparatively high electronegative nitrogen atoms, thereby leaving boron atoms with vacant p-orbitals. The intrinsic electronic configurations offer an astonishing bond tailoring opportunity, catalyzing robust crosslinking chemistry between rGO and hBNNS, thereby facilitating the formation of superior hybrid nanocomposite. Such interfacial chemistry consequentially decorates the surface with multi-functional adsorption sites that aid remarkable crosslinking chemistry with bio-conjugates. Crosslinking technique stands out among different immobilization techniques such as entrapment, physical adsorption, covalent binding, and sandwich incubation that elevate diffusion barrier, add up instability in protein interplay, and reduce denaturation in protein to enhance bio-activity, restrictions in conformational changes required during catalysis, the low adsorption efficiency of protein biomarkers and deteriorate charge transfer kinetics, etc. [15]. However, interfacial crosslinking actively promotes surface functionalization that specifically bridges the linking agent with adjacent protein biomarkers and offers high stability, capability to recycle with optimum catalytic efficiency, highly aligned protein conformation, and economical technique as well [15, 16].

The unique capability to engineer interfacial kinetics within OIHN significantly influences charge transfer kinetics. Weak (van der Waals, hydrogen bond, hydrophobic-hydrophilic) or strong (covalent or iono-covalent) interphase bonds are categorized into class I and class II, respectively [10]. These classifications greatly alter molecular dynamics

and sensing performance. Class II materials, resulting from stronger bonding chemistry between parent materials, offer enhanced stability in resulting nanocomposites, efficiently tune physical properties, enhance surface kinetics through proficient crosslinking of interfaces, and decorate multifunctional groups that enhance the loading capacity of protein biomarkers [10]. The synthesis of organic–inorganic nanocomposites plays a pivotal role, yet it presents significant challenges. Various existing methods often suffer from limitations such as the lack of scalability, complex preparation, and elevated temperatures [17–19]. Thus, in our study, we utilized a one-pot hydrothermal approach to simultaneously reduce GO, exfoliate hBNNS and rGO nanostructures, and reinforce crosslinking chemistry between them. This method operates under high pressure at low temperatures, reduces agglomeration, and prevents restacking of layers [20]. Optimized hydrothermal conditions and reducing agents enable simultaneous reduction of highly reactive GO nanosheets and effective bond tailoring, yielding the rGO-hBNNS hybrid nanocomposite.

In the present work, we utilize the synergy of OIHN interfacial chemistry to hydrothermally synthesize rGO-hBNNS hybrid nanocomposite endowed with exceptional electrocatalytic efficiency for label-free detection of CEA biomarker. The amine- and carboxyl-functional terminated groups on rGO decorated hBNNS nano-matrix offer effective site-specific anchoring by crosslinking Fc receptors of monoclonal Anti-CEA using EDC-NHS chemistry, thus preventing time-consuming, expensive, complex labeling procedures in clinical detection and commercial diagnostic kits. Furthermore, optimizing the immuno-interface environment improves binding affinity between rGO-hBNNS nano-matrix and Anti-CEA, thereby enhancing detection sensitivity and signal amplification. The superior electrocatalytic activity, coupled with selective anchoring of Anti-CEA in abundance onto nano-biointerface, augments precise conjugation between Anti-CEA and CEA, resulting in an exceptionally low detection limit of 5.47 pg/mL. Also, the proposed nano-biosensor based on rGO-hBNNS hybrid nanocomposite renders highly selective, stable, reproducible, and reliable outcomes for spiked serum sample analysis that comply with requirements for clinical diagnostics.

Experimental section

Chemicals and reagents

Various chemicals and reagents, including graphite powder (< 20 µm-particle size), bovine serum albumin (BSA), h-BN powder (98%, 5-µm particle size), 1-(3-(dimethylamino)-propyl)-3-ethyl carbodiimide hydrochloride (EDC), carcinoembryonic antigen (CEA), N-hydroxysuccinimide (NHS), and monoclonal antibodies of CEA, were procured from Sigma-Aldrich. Additionally, some chemicals and reagents were purchased from Thermo-Fisher Scientific and Merck, such as potassium permanganate ($KMnO_4$), sulfuric acid (H_2SO_4), hydrazine hydrate ($N_2H_4H_2O$, > 99%), isopropanol ($(CH_3)_2CHOH$), hydrogen peroxide solution (30% w/v H_2O_2), potassium ferrocyanide ($K_4[Fe(CN)_6]\cdot 3H_2O$), acetonitrile (CH_3CN), hydrochloric acid (HCl), sodium phosphate monobasic dihydrate ($NaH_2PO_4\cdot 2H_2O$), ethanol (C_2H_5OH), potassium ferricyanide ($K_3[Fe(CN)_6]$), and sodium phosphate dibasic dihydrate ($Na_2HPO_4\cdot 2H_2O$). All analytical grade reagents and chemicals were utilized, ensuring accuracy and precision. The human blood serum sample was obtained from All India Institute of Medical Sciences (AIIMS), New Delhi, with approval of the study. We utilized ultrapure Milli-Q water for solution preparation, sample washing. Fresh phosphate buffer saline (PBS) solution of different pH was prepared using solution A ($NaH_2PO_4\cdot 2H_2O$), solution B ($Na_2HPO_4\cdot 2H_2O$), and ultrapure Milli-Q water. All as-prepared solutions were stored at 4 °C until further use.

Instrumentation

Various techniques were used to systematically characterize hBNNS, GO, and rGO-hBNNS powder samples. The samples were analyzed to investigate their systematic phase using an X-ray diffractometer (XRD, Bruker Advance D-8). The morphology of the samples was monitored using scanning electron microscopy (SEM, EVO18 Zeiss) and transmission electron microscopy (TEM, TECNAI 200 kV). The chemical composition of the samples was determined using an energy dispersive X-ray spectrometer (EDX) and Fourier transform infrared spectroscopy (FTIR, Spectrum II-Perkin Elmer). Spectroscopic studies were carried out using a WITec Raman spectrometer ($\lambda = 532$ nm) and a UV–visible spectrometer (Perkin Elmer 950). Electrochemical Autolab Galvanostat/Potentiostat (Metrohm, Netherlands) was used to carry out differential pulse voltammetry (DPV) and cyclic voltammetry (CV) measurements. A three-electrode assembly comprising rGO-hBNNS–modified ITO substrates was employed as a working electrode, a platinum (Pt) rod as a counter electrode, and an Ag/AgCl rod in potassium chloride (KCl, 3 M) solution used as a reference electrode. The entire biosensing studies were carried out in ferro-ferricyanide ($[Fe(CN)_6]^{3-/4-}$, 5 mM) redox couple modified PBS solution (50 mM, 0.9% NaCl) of pH 7.0, which exhibits heterogeneous transfer of one electron.

One-pot hydrothermal synthesis rGO-hBNNS hybrid nanocomposite

As-prepared ultrathin multifunctional nanosheets of hBNNS synthesized via facile hydrothermal technique as described

in the previous literature [21]. hBNNS instead of bulk hBN was utilized for nanocomposite formation as partially ionic B-N bonds in bulk hBN limits dispersibility and exfoliation which thereby obstructs interfacial interactions with other precursors. Figure 1 depicts the preparation of nanocomposite, which started with dispersing 122 mg ultrathin nanosheets of hBNNS powder in 10 mL DI water in a beaker. Alongside, 90 mg of GO synthesized via eco-friendly improved Hummer's method (Electronic Supplementary Information, Fig. S1) was added to another beaker containing 10 mL DI water. Ultrasonication-assisted pre-treatment of both samples was carried alongside for 1 h at room temperature to boost up intercalation of hydroxyl ions that abet expansion of interlayer spacing while retaining their pristine crystalline structure. GO was relatively easily suspended in the solvent owing to its hydrophilic nature endowed by covalently functional groups that provide exceptional colloidal stability to them. Well-dispersed GO suspension was then added to the hBNNS solution, and the resulting mixture was again kept for ultrasonication for 1 h to assist reinforcement of intercalated precursors. Then, 1 mL hydrazine hydrate and 15 mL DI water were added to the resultant GO-hBNNS dispersion followed by continuous stirring of 270 rpm for 6 h at room temperature. The homogeneous and stable dispersion was poured into 50 mL Teflon and placed in a stainless steel autoclave vessel. Next, the autoclave vessel was placed inside a high-temperature oven for 12 h at 180 °C to obtain rGO-hBNNS nanocomposite. Interestingly, electron-deficient boron sites (Lewis acid) capped with Lewis base (amine) established stable bond affinity towards electrophilic GO nanosheets that led to the successful synthesis of rGO-hBNNS nanocomposite. The resultant rGO-hBNNS hybrid nanocomposite was washed to remove ionic impurities present with DI water via centrifugation of 7000 rpm until pH 7.0 was attained. The nanocomposite was dried in a vacuum oven at 85 °C for 24 h.

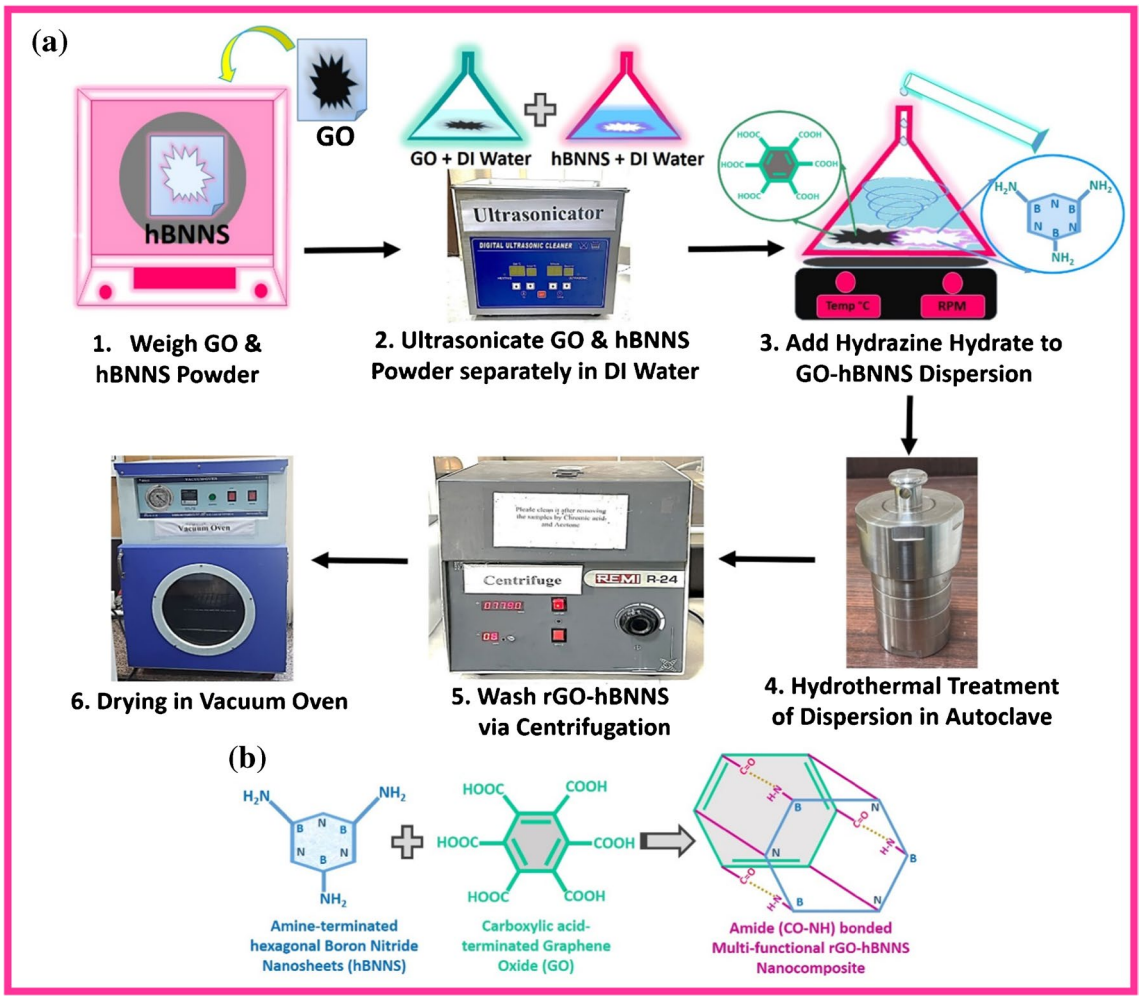


Fig. 1 **a** Steps involved in one-pot hydrothermal synthesis of rGO-hBNNS hybrid nanocomposite and **b** interfacial chemistry between GO and hBNNS for nanocomposite formation

Fabrication of rGO-hBNNS film onto ITO and nano-biosensing platform

The electrophoretic deposition (EPD) technique was employed for fabricating rGO-hBNNS/ITO electrodes utilizing the synergy of dispersion and voltage drifting (Fig. 2a). We achieved the formation of a uniform colloidal suspension containing 0.25 mg/mL of rGO-hBNNS nanocomposite in acetonitrile (CH_3CN) by adding 300 µL of magnesium nitrate ($MgNO_3$, 1 mg/mL). Suspension stability was enhanced through continuous ultrasonication for 1 h. The uniform suspension is poured into a beaker comprising a two-electrode assembly of pre-hydrolyzed ITO (0.8 cm × 2.5 cm) as the working electrode at the positive terminal and platinum rod as the counter electrode at the negative terminal. ITO electrodes were hydrolyzed prior to EPD with hydrogen peroxide, ammonia, and DI water in 1:1:5 ratios. This pre-treatment of ITO substrates helps in activating the radical groups on the surface. In the following step, a homogeneous film of rGO-hBNNS (dimension, 0.8 cm × 0.8 cm) was deposited onto the ITO electrode by applying an external DC voltage of 15 V for 1 min.

For fabricating an optimized nano-biosensing platform, a stock solution of monoclonal Anti-CEA was prepared in PBS solution (pH 7.4) by activating –COOH, $-NH_2$ groups of antibody utilizing EDC-NHS chemistry (Fig. 2b). Briefly, Anti-CEA (50 µg/mL) was conjugated with EDC (0.2 M) and NHS (0.05 M) in a 2:1:1 ratio at room temperature for 30 min. EDC being a zero-length cross-linker reacts with carboxyl terminated Fc region (L-glutamic acid) of Anti-CEA [22] to form a reactive as well as unstable O-acylisourea ester that becomes stabilized with amine-reactive NHS ester. The EDC-NHS intermediate crosslinks the carboxyl-terminated Fc region of Anti-CEA with an amine-terminated hBNNS interface to form an amide bond (Fig. 3a) [23], while lysine residue on

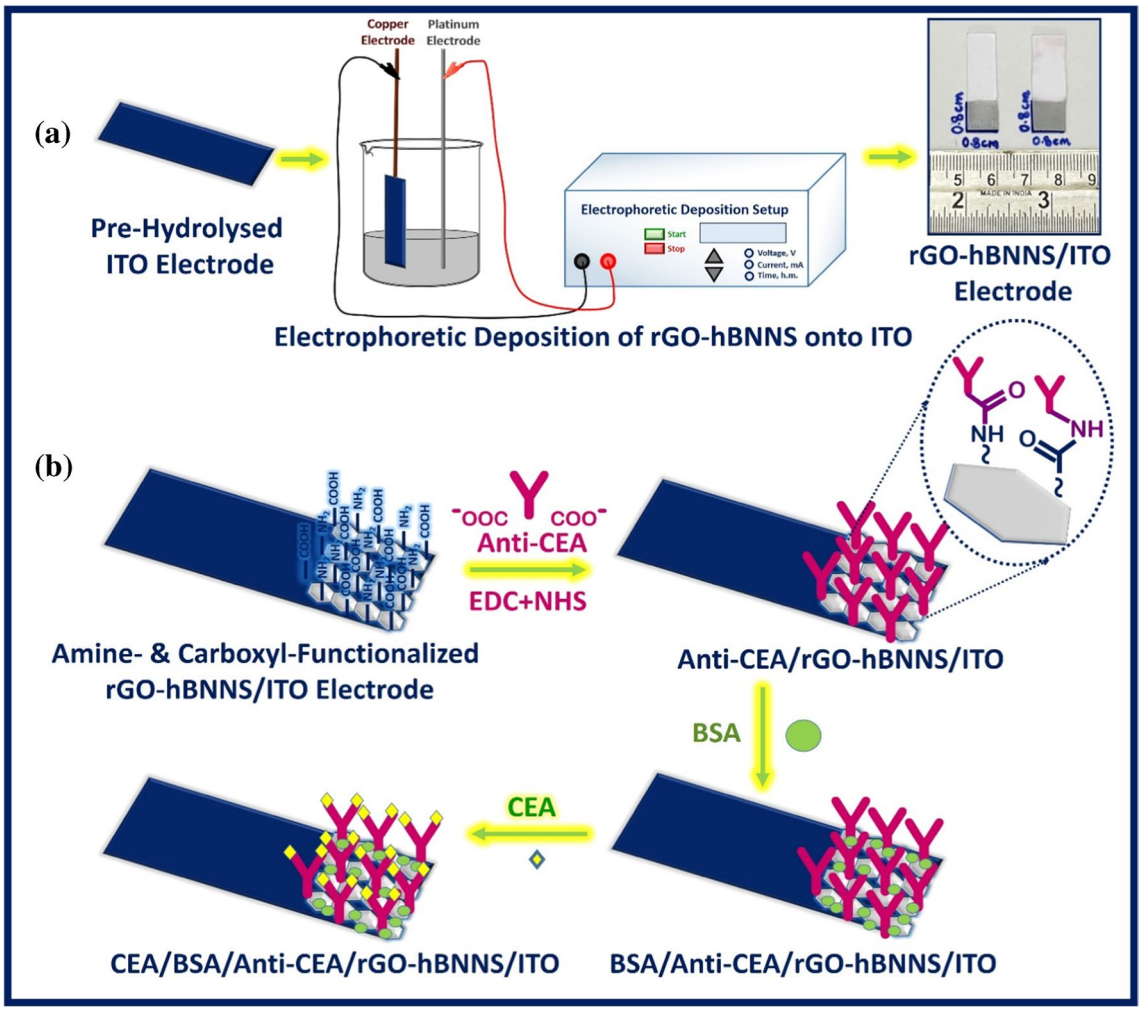


Fig. 2 Schematic diagram on **a** fabrication of rGO-hBNNS/ITO electrode by EPD; **b** modification of rGO-hBNNS/ITO electrode by Anti-CEA, BSA, and CEA to fabricate nano-biosensor

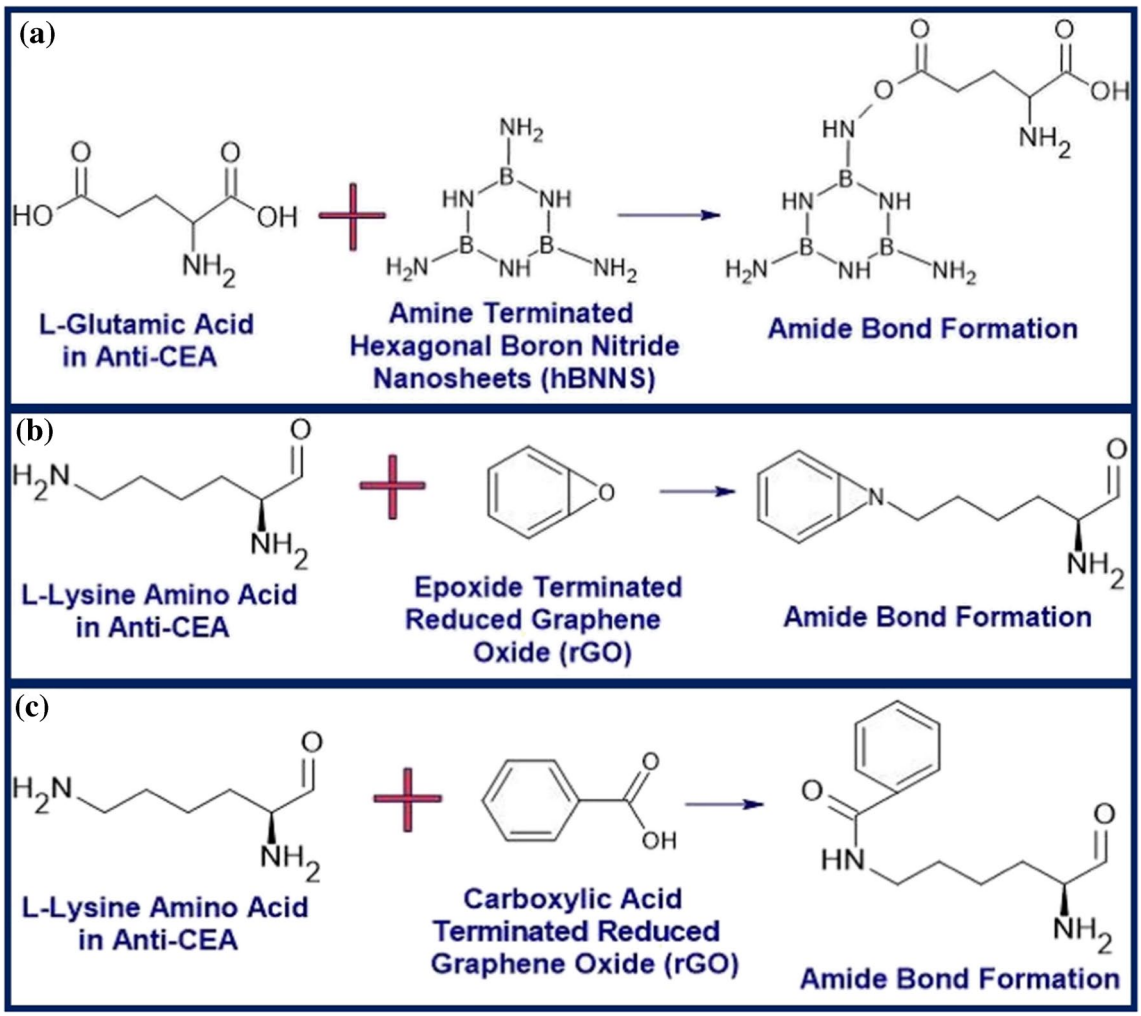


Fig. 3 Amide bond formation between **a** L-glutamic acid and amine-terminated hBNNS, **b** L-lysine and epoxide terminated rGO, **c** L-lysine and carboxylic acid terminated rGO, respectively

Anti-CEA offers amino-groups in side-chains that conjugate to C–O–C and–COOH terminated functional groups on rGO interface via EDC-NHS coupling (Fig. 3b–c) [24]. Then, 20 µL of Anti-CEA was immobilized onto uniform rGO-hBNNS film for 24 h in a humid chamber for optimized immobilization. Figure 2b depicts steps involved in the fabrication of nano-biosensor that include crosslinking of Anti-CEA with rGO-hBNNS/ITO nano-biointerface and CEA conjugation with Anti-CEA. On Anti-CEA/rGO-hBNNS/ITO immunoelectrode, 10 µL BSA (1% w/w in ultrapure Milli-Q water) was added to block the non-specifically adsorbed Anti-CEA sites. In the final step, 10 µL CEA was added and incubated for 30 min onto BSA/Anti-CEA/rGO-hBNNS/ITO immunoelectrode of required concentration. The as-prepared immunoelectrodes were stored at 4 ℃ until further use.

Results and discussion

Material characterization

X-ray diffraction

A systematic XRD investigation was carried out in the range of 5°–80° to study phase formation, phase purity, and crystallinity of as-synthesized (I) hBNNS, (II) GO, and (III) rGO-hBNNS nanocomposite as shown in Fig. 4a. Optimized hydrothermal conditions and efficient intercalation-exfoliation chemistry of IPA and hydrazine hydrate yielded a pure hexagonal phase of BN corroborated by two diffraction peaks at 26.84° and 41.72° as shown in Fig. 4a(I) that index to (002) and (100) crystallographic planes, respectively [JCPDS Card No. 034–0421] [21].

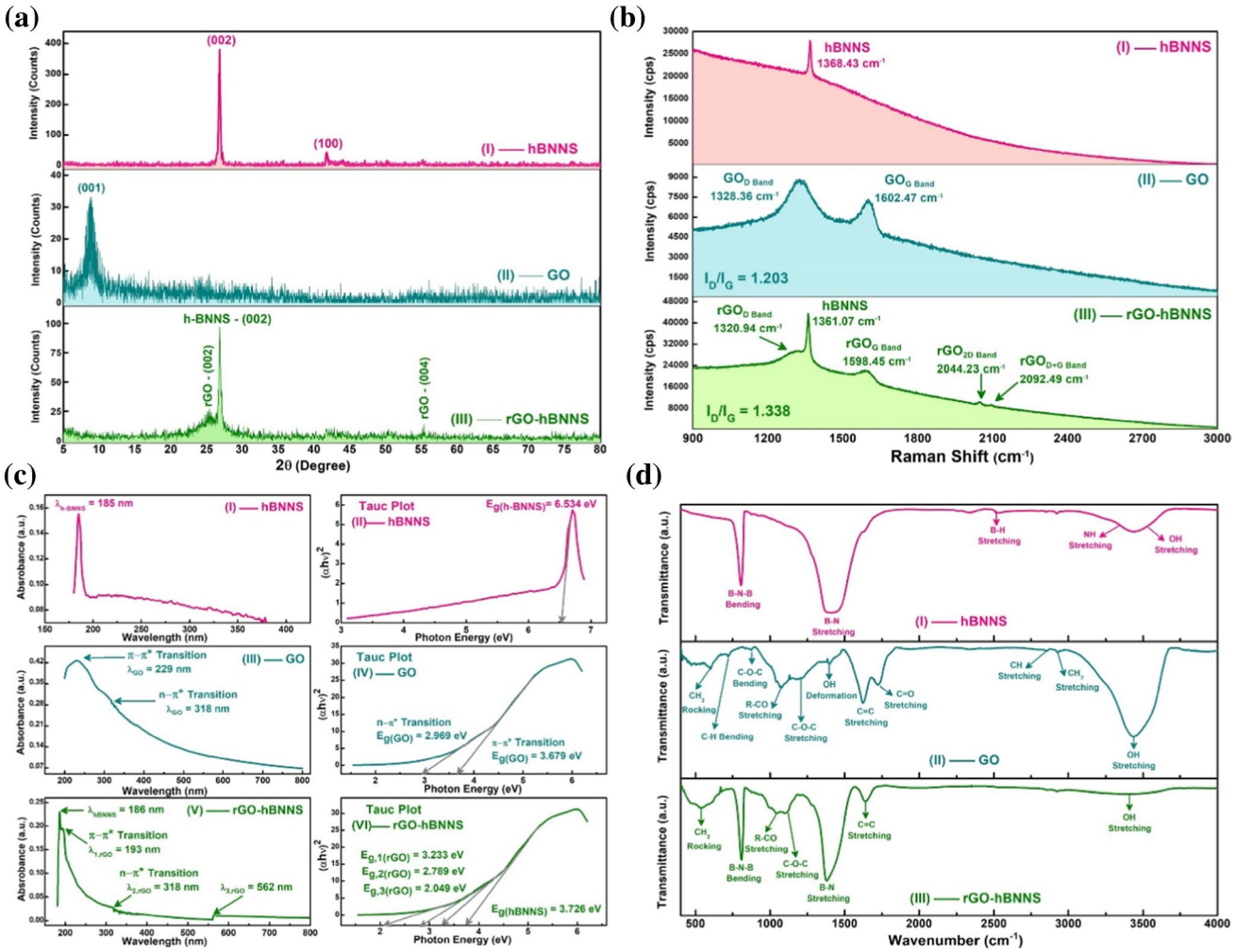


Fig. 4 **a** XRD, **b** Raman, **c** UV–visible, and **d** FT-IR spectroscopy of (I) hBNNS, (II) GO, and (III) rGO-hBNNS samples, respectively

A distinct, intense, broad peak associated with GO significantly down-shifted to a lower diffraction angle at $2\theta = 8.8°$ that corresponds to (001) characteristic plane (Table 1) indicates exceptional oxidation of bulk graphite powder under the influence of harsh precursors comprising $KMnO_4$ and H_2SO_4 as shown in Fig. 4a(II) [25]. A significant down-shift in diffraction peak and large d-spacing ($d_{(001)}$ = 1.004 nm) evaluated using Bragg's relation (Eq. 1) represents characteristic changes GO undergone due to complete intercalation of hydroxyl, carboxyl, carbonyl, and epoxy groups. The intercalation facilitated facile exfoliation of π-bonded bulk graphite consequently boosting oxidation of GO sheets, verified by FT-IR also. The XRD pattern of rGO-hBNNS nanocomposite in Fig. 4a(III) depicted an intense diffraction peak at 26.84° that accord with (002) crystallographic plane of hBNNS, whereas two

Table 1 Intense hkl plane, peak position, FWHM, crystallite size, and d-spacing associated with samples hBNNS, GO, and rGO-hBNNS nanocomposite, respectively

S.No.	Sample		Intense (hkl) planes	Peak position, 2θ (°)	FWHM (°)	Crystallite size (D, nm)	d-spacing (nm)
1	Hexagonal boron nitride nanosheets, hBNNS		(002)	26.84	0.2392	34.171	0.332
2	Graphene oxide, GO		(100)	8.8	1.205	6.616	1.004
3	rGO-hBNNS Nanocomposite	hBNNS	(002)	26.84	0.3187	25.640	0.332
		rGO	(002)	25.4	1.089	7.483	0.350

relatively diminished peaks appeared at 25.4° and 55.3° befits (002) and (004) diffraction planes of rGO indicting preferential growth of relatively poor crystallinity rGO over highly crystalline hBNNS nanostructures by hydrothermal synthesis.

$$Bragg's relation : 2d \times Sin\theta = n\lambda \quad (1)$$

where d stands for interplanar spacing (in nm), θ is Bragg's diffraction angle (in radians), n is the integer for the order of planes, and λ is the wavelength of incoming electron.

Interestingly, the appearance of two diffraction planes in rGO and a profound decrease in d-spacing was observed on substantial hydrothermal reduction of GO ($d_{(100)} = 1.004$ nm) to rGO ($d_{(002)} = 0.350$ nm) on nanocomposite formation that explains disordered stacking of layered nanostructures on restoration of graphitic structure in rGO. In addition, hBNNS nanostructures in nanocomposite attained highly crystalline morphology in crystallographic layers that preferentially oriented in only (002) hkl plane while eliminating (100) crystallographic planes. Moreover, intercalation, growth, and nucleation of rGO nanosheets under in situ hydrothermal constraints improved the quantum confinement effect in rGO-hBNNS nanocomposite by remarkably reducing crystallite size estimated using Debye–Scherrer's formula (Eq. 2) from 34.17 nm in pure hBNNS to 25.64 nm in rGO-hBNNS nanocomposite while retaining the d-spacing of hBNNS (Table 1). These findings indicate the formation of nano-sized flakes comprising fewer layers of hBNNS in the nanocomposite. On the contrary, the crystallite size of rGO increased to 7.483 nm from 6.616 nm in GO which ascribes to increased stacking in crystallographic planes of rGO to accommodate the difference in d-spacing between hBNNS and rGO for attaining stable nano-confined geometry (Table 1).

$$Debye - Scherrer's formula : D = \frac{0.9\lambda}{\beta \times Cos\theta} \quad (2)$$

where D stands for average crystallite size (in nm), λ is the wavelength of incoming X-rays (in nm), β represents full-width half maxima (in radians), and θ is the diffraction angle (in radians). Furthermore, the d-spacing of rGO ($d_{(002)} = 0.350$ nm) was found consistent with the d-spacing of graphite ($d_{(002)} = 0.350$ nm), indicating restoration of graphitic domains in basal planes and adsorption of hydrophilic oxygen moieties on the surface instead of interplanar domains [25].

Raman spectroscopy

Raman spectroscopy is a powerful non-destructive tool utilized for investigating the electronic structure, nanocomposite formation, and ordering of layers by analyzing the appearance of vibrational modes, changes in peak intensity, peak shift, FWHM, and defect density in as-synthesized samples (I) hBNNS, (II) GO, and (III) rGO-hBNNS (Fig. 4b). The appearance of a prominent peak at 1368.43 cm^{-1} in Fig. 4b(I) demonstrates the characteristic in-plane E_{2g} phonon mode of hBNNS. Figure 4b(II) depicts typical Raman peaks in GO at around 1328.36 cm^{-1} and 1602.47 cm^{-1}, which represents the first-order breathing mode of k-point phonons of A_{1g} symmetry due to sp^3 hybridized defect-induced states (D-band) and sp^2 hybridized E_{2g} phonon mode (G-band), respectively [26]. However, an overtone of D-band denoted by 2D band (or G'-band) in GO was absent indicating the formation of highly defect-induced nanosheets due to the lack of ordered stacking along the c-axis [27]. Formation of rGO-hBNNS nanocomposite was marked by three peculiarly red-shift peaks as shown in Fig. 4b(III) that corresponds to the D-band of rGO, in-plane E_{2g} vibrational mode of hBNNS and G-band of rGO at 1320.94 cm^{-1}, 1361.07 cm^{-1}, and 1598.41 cm^{-1}, respectively. Interestingly, two diminished peaks at 2044.23 cm^{-1} and 2092.49 cm^{-1} originated in rGO-hBNNS that corresponds to immensely red-shifted 2D and D + G bands, respectively, implying restoration of ordered stacking in rGO nanosheets directed in prominent (002) orientation of hBNNS along c-axis [27, 28]. A significant red shift in the 2D band of rGO-hBNNS is attributed to the transfer of electron-cloud density from rGO to functionalized hBNNS leading to stable bond alignment in defect-induced domains [29].

The prominent red-shift phenomenon in rGO of about 7.42 cm^{-1} (D-band) and 4.06 cm^{-1} (G-band), respectively, explains the complete reduction of GO to rGO by restoration of graphitic sp^2 hybridized nanostructures, increase in electron density, and successful interface coupling between rGO and hBNNS under influence of exceptional tensile strain on nanocomposite formation [30]. However, a red-shift of 7.36 cm^{-1} in hBNNS on nanocomposite formation indicates synthesis of improved crystalline nanostructures of hBNNS, a remarkable reduction in layer thickness, and expansion in the lateral size of nanosheet morphology owing to in-built tensile strain developed amid hydrothermal treatment. The noteworthy increase in I_D/I_G ratio on reduction of GO (1.203) to rGO (1.338) amid nanocomposite synthesis indicates an increase in layered stacking in rGO due to high-temperature reduction that induced unrepaired defects in rGO [25] and verifies the increased crystallite size in XRD. Additionally, elevated I_D/I_G ratio supports red-shift phenomena observed in Raman peaks on nanocomposite synthesis attributed to the consequence of the effective reduction of oxygen functional moieties in GO that materialize disorder, vacancies, distortion, and defects in sp^2 carbon domains for creating a stable interface between rGO and hBNNS molecules. Interestingly, the I_{2D}/I_G ratio appreciably increased to 0.472 indicating restoration of crystallinity and reduction

of layer stacking in as-synthesized rGO-hBNNS nanocomposite [31]. Furthermore, the absence of any extra peaks in the Raman spectra of nanocomposite indicates exceptional phase purity in as-synthesized samples.

UV–visible spectroscopy

Investigating spectral transmittance profiles via UV–visible spectroscopy before and after nanocomposite formation allowed us to gain more understanding of interaction and transformation that resulted in bandgap tuning and absorption peaks shifting as depicted in Fig. 4c. A red-shift was observed in the sharp absorption edge associated with hBNNS from 185 nm (6.534 eV) in pure hBNNS to 186 nm (3.726 eV) on formation of rGO-hBNNS nanocomposite (Fig. 4c(I, II, V, VI)). From UV–visible spectra of highly disordered GO nanosheets (Fig. 4c(III, IV)), two characteristic absorption peaks positioned at 229 nm (3.679 eV) and 316 nm (2.969 eV) that correspond to π-π^* transitions associated with C = C bonding and n-π^* transitions associated with C = O bonding are observed, respectively. On hydrothermal reduction of GO, the absorption peaks related to π-π^* transition blue-shifted to 193 nm (3.233 eV), while n-π^* transition retained absorbance peak to 318 nm (2.789 eV), and a new peak emerged at 562 nm (2.049 eV), respectively (Fig. 4c(V, VI)) [32]. The bandgap of as-synthesized samples was estimated by extrapolating the linear region in Tauc's plot demonstrated between $(\alpha h\upsilon)^2$ versus $(h\upsilon)$ [14]. An exceptional bandgap tuning from 5.34 eV in pure hBNNS to 3.726 eV and from 3.679 eV in GO to 3.233 eV after rGO-hBNNS nanocomposite formation was observed owing to the removal of hydrophilic oxygen moieties and layer-layer interaction between rGO and hBN nanosheets that induced structural changes in the nanocomposite that lead to dispersion in electronic band as shown in Fig. 4c(II, IV, VI), respectively.

Fourier transform infrared spectroscopy

Analysis of chemical bond formation and functionalization in as-synthesized samples was investigated by FT-IR as shown in Fig. 4d. FT-IR transmittance peaks associated with hBNNS (Fig. 4d(I)) described in Table S1 (Electronic Supplementary Information) demonstrate fingerprint peaks positioned at 804 cm^{-1} and 1413 cm^{-1} in addition to functionalized peaks at 2500, 3200–3500, and > 3500 that corresponds to B-N-B bending, B-N stretching, B-H stretching, N–H stretching, and O–H stretching, respectively [21]. The characteristic skeletal vibration at 1621 cm^{-1} is attributed to the retention of sp^2 hybridized graphitic domains that form the backbone in GO (Fig. 4d(II)) [33]. The bonding configuration of GO was found to exhibit hydrophilic oxygen peaks of epoxy (874 cm^{-1}, 1216 cm^{-1}), alkoxy (1074 cm^{-1}), carboxyl (1729 cm^{-1}), and hydroxyl (1398 cm^{-1}, 3437 cm^{-1}) groups, respectively (Table S1) [33]. Additional functionalized peaks that belong to CH_2 rocking (603 cm^{-1}), C-H out of plane bending (819 cm^{-1}), alkyl (C-H) symmetric (2832 cm^{-1}), and asymmetric (2855 cm^{-1}) stretch emerged in GO [34]. Figure 4d(III) depicts the material-specific peak of (1) hBNNS that include B-N-B bending (808 cm^{-1}) and B-N stretching (1379 cm^{-1}) and (2) rGO that comprises skeletal vibration of graphitic domain C = C (1637 cm^{-1}) that attribute to successful formation rGO-hBNNS nanocomposite. Yet, some alkyl (546 cm^{-1}) and oxygen-containing moieties (1036 cm^{-1}, 1079 cm^{-1}, 3392 cm^{-1}) still appeared after exceptional hydrothermal reduction of GO, indicating a partial reduction of GO nanosheets (Fig. 4d(III)).

Morphological and elemental analysis

The micro- and nano-structure morphology together with the elemental composition of as-synthesized (I) hBNNS, (II) GO, and (III) rGO-hBNNS was investigated using SEM, TEM, and EDAX techniques as shown in Fig. 5a, b, and c, respectively. The SEM micrograph depicted in Fig. 5a(I) demonstrates well-nucleated enlarged flaky nanosheets of hBN, while the surface morphology of GO displays corrugated edges of heaped-up nanosheets, confirming the "extensive peeling" of bulk graphite powder (Fig. 5a(II)). These curls and folds in edges indicate the formation of highly flexible GO nanosheets. A blend of several dark and bright regions in rGO-hBNNS nanocomposite in Fig. 5a(III) depicts the conglomeration of wavy wrinkled rGO nanosheets with relatively sleek disc-shaped hBNNS, respectively. This aggregation in rGO and hBN nanosheets materializes to minimize their surface energy cropped up by the presence of dangling bonds at the surface. Figure 5b(I) displays a more comprehensive view of hBNNS exhibiting transparent circular nanosheets attributing to the formation of ultrathin nanosheets. The TEM image in Fig. 5b(II) displays wrinkled, transparent 2D nanosheets of GO indicating oxygen-intercalated corrugated interlayers. Figure 5b(III) displays intertwined regions of rGO nanosheets with hBNNS, thereby confirming the successful reinforcement to facilitate the formation of rGO-hBNNS hybrid nanocomposite.

The corresponding elemental data (Table S2, Electronic Supplementary Information) obtained from EDX as shown in Fig. 5c(I–III) provides clarity on the formation of pure hBNNS, GO, and rGO-hBNNS. The increment of the C/O atomic ratio from 1.997 in GO to 7.966 in rGO on nanocomposite synthesis indicates a remarkable reduction of oxygen moieties. A significant C/N ratio (0.371) and relatively enhanced atomic % and weight % of nitrogen in rGO-hBNNS ascribe towards hydrazine hydrate-assisted reduction of GO that lead to enhanced nitrogen-mediated

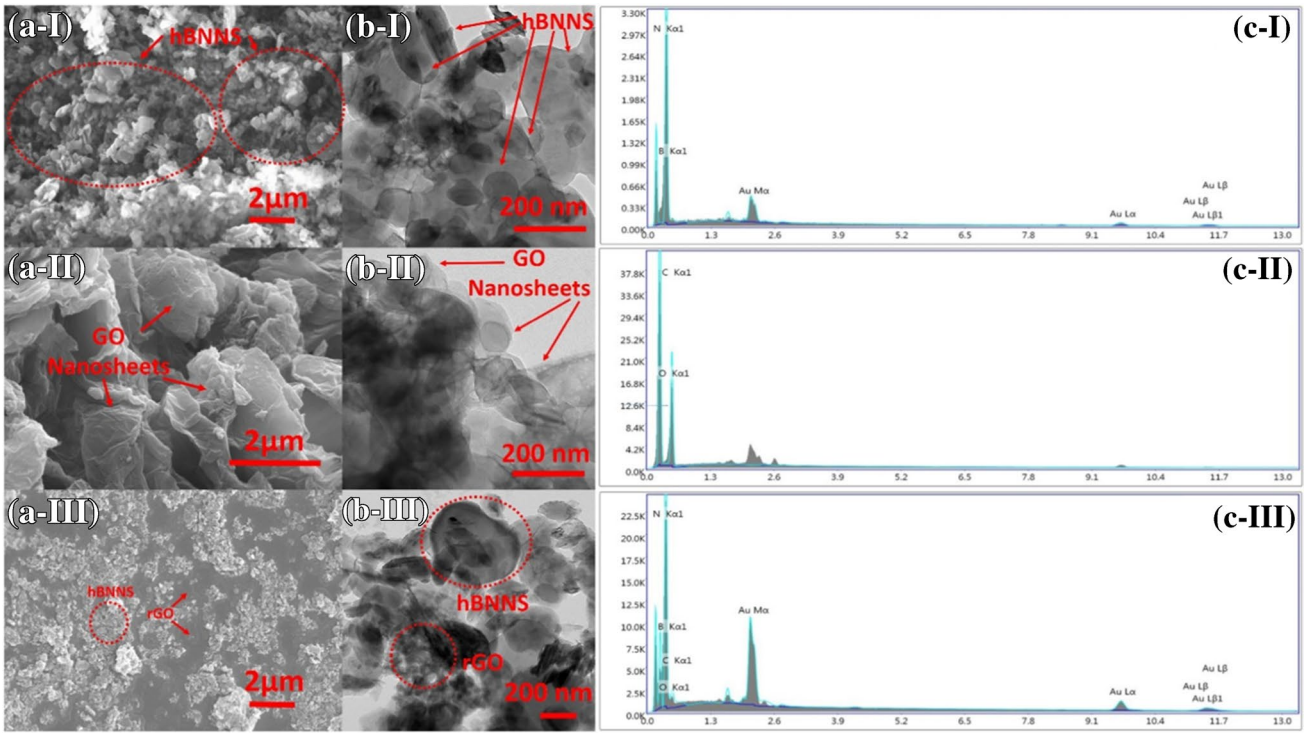


Fig. 5 **a** SEM (2 µm), **a** TEM (200 nm), and **c** EDAX of (I) hBNNS, (II) GO, and (III) rGO-hBNNS nanocomposite, respectively

crosslinking of bond between rGO and hBNNS that significantly promoted nanocomposite formation.

Electrochemical characterization

Electrode study on immunosensor fabrication

To investigate the optimized time required for antibody binding, FT-IR spectra of Anti-CEA were monitored after immobilization time of 12 h (blue) and 24 h (purple) onto rGO-hBNNS/ITO electrodes, respectively (Fig. 6a). FT-IR spectra before and after Anti-CEA immobilization exhibited fingerprint peaks assigned to B-N-B bend, B-N stretch, C = C stretch, alkyl rock, alkoxy stretch, and epoxy stretch group formerly possessed by rGO-hBNNS nanocomposite, while protein-specific peaks in regions II, III, and IV appeared on Anti-CEA immobilization that corresponds to protein phosphorylation (869 cm^{-1}), amide III (NH bend, 1264 cm^{-1}), amide II (CN stretch, 1567 cm^{-1}), and amide I (C = O stretch, 1647 cm^{-1}), respectively [21]. After an incubation time of 24 h, prominent and unique peaks emerged that assigned to EDC-NHS activation (2081 cm^{-1}), aliphatic stretching (2729 cm^{-1}, 2877 cm^{-1}), amide B stretching (2978 cm^{-1}), amide A stretching (3300 cm^{-1}), and OH stretching (> 3500 cm^{-1}) [21, 35], indicating complete immobilization of Anti-CEA. Interestingly, Anti-CEA immobilized for 12 h displayed protein-specific peaks, but the above-mentioned unique peaks were absent, indicating partial immobilization of antibodies after 12 h.

Subsequently, FT-IR investigated immunosensor fabrication on successive protein immobilization onto rGO-hBNNS nanocomposite (Fig. 6b). Immobilization of Anti-CEA modified with EDC-NHS was evident by the appearance of amide bands of protein structure in addition to the characteristic peaks of h-BNNS in Anti-CEA/h-BNNS as discussed above [36]. Following antibody immobilization, transmittance intensity of FT-IR spectra reduced significantly in regions II, III, IV, and V owing to BSA (10 µL) immobilization onto the Anti-CEA/rGO-hBNNS matrix, implying effective passivation of non-specifically adsorbed antibodies [21]. CEA was immobilized onto BSA/Anti-CEA/rGO-hBNNS immunoelectrode that led to a transmittance intensity increase, and shifted region III, IV, V, and VI on addition of CEA onto BSA-modified platform depict successful crosslinking of antibody-antigen complexes [37].

Electrochemical characterizations were performed including CV (Fig. 6c) and DPV (Fig. 6d) performed after each step of immunosensor fabrication that supports FT-IR outcome. Prior to electrochemical investigations, we actively subjected the as-fabricated Anti-CEA/rGO-hBNNS/ITO electrode to a pH study (Electronic Supplementary Information, Fig. S2). This proactive approach ensured meticulous optimization of the sensing environment, resulting in robust and stable measurements. The scan rate study (Electronic

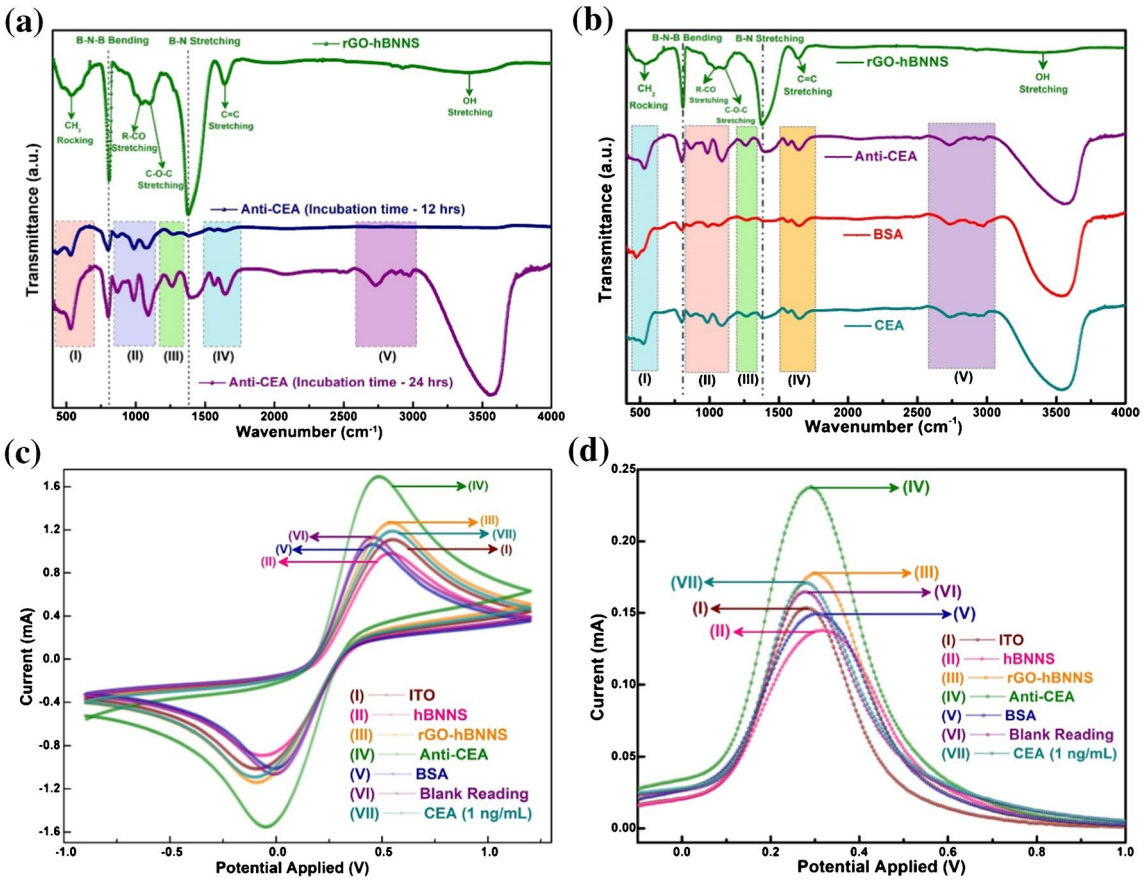


Fig. 6 **a** Optimization of Anti-CEA immobilization time by FT-IR spectra and fabrication of rGO-hBNNS-based immunoelectrode monitored by **b** FT-IR, **c** CV, and **d** DPV, respectively

Supplementary Information, Fig. S3) conducted on the BSA/Anti-CEA/rGO-hBNNS/ITO electrode compellingly demonstrated diffusion-controlled electron transfer dynamics at the interface immunoelectrode. This comprehensive analysis provided insightful evidence of the electrode's superior performance in facilitating efficient electron transfer processes, underscoring its suitability for sensitive and accurate detection applications. Further, we investigated CV study on successive protein immobilization, and the data obtained was utilized to determine diffusion coefficients (Table 2) after the sequential protein modification step by the Randles–Sevcik equation (Eq. 3) to study the interfacial kinetics in preparation of immunosensor [38].

$$I_p = \left(2.69 \times 10^5\right) n^{3/2} A C D^{1/2} \sqrt{\nu} \tag{3}$$

where I_p is the peak (anodic or cathodic) current, n is the number of electrons transferred during the redox reaction ($n = 1$), A denotes the electroactive surface area of the electrode (0.64 cm^2), C stands for the concentration of the

Table 2 Depiction of anodic peak current obtained from CV, evaluation of diffusion coefficient, and DPV current on fabrication of immunoelectrode

Stepwise electrode modification	Anodic peak current (I_{pa}, mA)	Diffusion coefficient (D, $cm^2\ s^{-1}$)	DPV current (mA)
ITO	$1.11 \pm (1.0 \times 10^{-2})$	$3.32 \times 10^{-7} \pm (6.1 \times 10^{-9})$	$0.154 \pm (1.1 \times 10^{-3})$
hBNNS	$0.964 \pm (1.6 \times 10^{-2})$	$2.51 \times 10^{-7} \pm (8.6 \times 10^{-9})$	$0.138 \pm (1.1 \times 10^{-4})$
rGO-hBNNS	$1.25 \pm (1.3 \times 10^{-2})$	$4.23 \times 10^{-7} \pm (8.6 \times 10^{-9})$	$0.177 \pm (3.7 \times 10^{-4})$
Anti-CEA	$1.67 \pm (1.9 \times 10^{-2})$	$7.55 \times 10^{-7} \pm (1.7 \times 10^{-8})$	$0.238 \pm (4.6 \times 10^{-4})$
BSA	$1.07 \pm (0.15 \times 10^{-2})$	$3.07 \times 10^{-7} \pm (8.9 \times 10^{-10})$	$0.149 \pm (1.8 \times 10^{-4})$
Blank reading	$1.13 \pm (0.15 \times 10^{-2})$	$3.44 \times 10^{-7} \pm (9.4 \times 10^{-10})$	$0.165 \pm (9.2 \times 10^{-5})$
CEA	$1.19 \pm (0.08 \times 10^{-2})$	$3.81 \times 10^{-7} \pm (5.2 \times 10^{-10})$	$0.171 \pm (5.5 \times 10^{-4})$

redox species (5×10^{-3} mol cm^{-3}), D is the diffusion coefficient, and ν is the scan rate (50 mV/s). Upon electrophoretic deposition (EPD) of rGO-hBNNS onto ITO, we observed anodic peak currents, DPV currents, and diffusion coefficients of $1.25\pm(1.3\times10^{-2})$ mA, $0.177\pm(3.7\times10^{-4})$ mA, and $4.23\times10^{-7}\pm(8.6\times10^{-9})$ cm^2 s^{-1}, respectively. This deposition resulted in a significant enhancement of approximately ~ 29%, ~ 29%, and ~ 68%, respectively, compared to the deposition of h-BNNS onto ITO, which showed anodic peak currents, DPV currents, and diffusion coefficients of $0.964\pm(1.6\times10^{-2})$ mA, $0.138\pm(1.1\times10^{-4})$ mA, and $2.51\times10^{-7}\pm(8.6\times10^{-9})$ cm^2 s^{-1}, respectively (Table 2).

The observed enhancements indicate the superior electrochemical performance of the electrophoretically deposited rGO-hBNNS film on ITO compared to h-BNNS deposition. This outcome indicates the key role of electron-rich rGO nanosheets within the rGO-hBNNS nanocomposite, compared to bare hBNNS, in enhancing electrocatalytic activity by promoting exceptional electron transfer kinetics. The incorporation of rGO-hBNNS leads to increased electrochemical activity and improved mass transport properties, thereby establishing it as a highly promising material for electrochemical sensing applications. Following the immobilization of Anti-CEA, we observed a substantial increase in the anodic peak current and DPV current, approximately $1.67\pm(1.9\times10^{-2})$ mA and $0.238\pm(4.6\times10^{-4})$ mA, respectively, indicating enhancements of approximately ~ 33% and ~ 34% compared to rGO-hBNNS/ITO electrode. Additionally, there was a significant improvement in the diffusion coefficient to $7.55\times10^{-7}\pm(1.7\times10^{-8})$ cm^2 s^{-1}, representing a nearly ~ 78% enhancement compared to rGO-hBNNS/ITO electrode. The substantial electron kinetics on Anti-CEA binding attribute to increased loading capacity offered by multifunctional (-COOH, $-NH_2$) nano-biointerface sites on rGO-hBNNS mediated with EDC-NHS coupling chemistry.

Upon the addition of globular protein (BSA) onto the Anti-CEA/rGO-hBNNS/ITO immunoelectrode, we observed a significant reduction in the anodic peak current, DPV current, and diffusion coefficient, measuring approximately $1.07\pm(0.15\times10^{-2})$ mA, $0.149\pm(1.8\times10^{-4})$ mA, and $3.07\times10^{-7}\pm(8.9\times10^{-10})$ cm^2 s^{-1}, respectively. This reduction serves as clear evidence for the successful passivation of disoriented antibodies, indicating improved electrode stability and specificity in the detection of CEA. A significant increase in anodic peak current, DPV current, and diffusion coefficient was observed for the blank reading (0 ng/mL) compared to the BSA/Anti-CEA/rGO-hBNNS/ITO immunoelectrode, measuring $1.13\pm(0.15\times10^{-2})$ mA, $0.165\pm(9.2\times10^{-5})$ mA, and $3.44\times10^{-7}\pm(9.4\times10^{-10})$ cm^2 s^{-1}, respectively. The rise in current measurements suggests that ions in the phosphate buffer solution (pH-7.4) aid electron transfer processes, leading to increased current responses in electrochemical measurements. This highlights the vital role of the buffer solution in stabilizing the electrochemical environment and improving measurement sensitivity. Following the subsequent immobilization of CEA, a further increase in the anodic peak current, DPV current, and diffusion coefficient was observed, measuring $1.19\pm(0.08\times10^{-2})$ mA and $0.171\pm(5.5\times10^{-4})$ mA and $3.81\times10^{-7}\pm(5.2\times10^{-10})$ cm^2 s^{-1}, respectively. This increase confirms the successful formation of the antibody-antigen complex, providing clear and unequivocal evidence of specific binding between the immobilized antibodies and the target antigen, thereby affirming the effectiveness of the BSA/Anti-CEA/rGO-hBNNS/ITO detection platform [21].

Response study

Figure 7a depicts the calibration curve monitored using the DPV technique on immobilization of successive

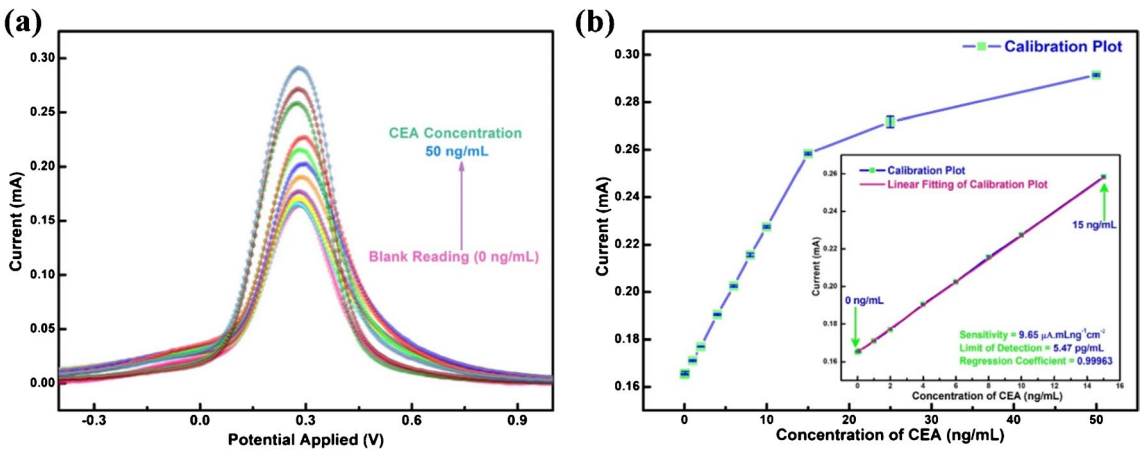


Fig. 7 **a** DPV response on calibrating CEA concentration from 0 to 50 ng/mL and **b** calibration graph between peak DPV current and CEA concentration

concentrations of carcinoembryonic antigen in the physiological range from 0 to 50 ng/mL owing to elevation of CEA beyond 5 ng/mL in NSCLC patients. The DPV current increased in the presence of elevated concentrations of CEA w.r.t. blank reading up to 15 ng/mL indicating linearity in current response owing to proportional electron transfer promoted by enhanced surface active sites of rGO-hBNNS for antibody-antigen mediation on the nano-biointerface (Fig. 7b). However, the linear response in peak current starts to deviate beyond 15 ng/mL ascribing the saturation of crosslinking sites on nano-biointerface (BSA/Anti-CEA/rGO-hBNNS/ITO) due to optimized occupancy of surface functional moieties.

Fitting of a linear plot depicting the calibration curve in Fig. 7b assisted computation of key parameters for assessing the sensing performance of as-fabricated nano-biosensor based on rGO-hBNNS/ITO electrode. The linear range fitted from 0 to 15 ng/mL with regression equation of form $y = mx + c$ having regression coefficient (R^2) of 0.99963 is depicted in Eq. (4).

$$I(mA) = [9.65\mu AmLng^{-1}cm^{-2} \times (CEAconcentration) + 0.165mA] \quad (4)$$

The slope associated with linear relationship aided the calculation of sensitivity (9.65 µA mL ng^{-1} cm^{-2}) attained by as-fabricated immunosensor based on rGO-hBNNS by utilizing slope/surface area (0.64 cm^2) formula. Further, the standard deviation in nano-biointerface (BSA/Anti-CEA/rGO-hBNNS/ITO) together with the sensitivity of the sensor helped in determining the lower limit of detection [39] using Eq. (5),

$$\text{Limit of detection} = 3\sigma/m \quad (5)$$

where σ denotes the standard deviation of blank immunoelectrode (BSA/Anti-CEA/rGO-hBNNS/ITO) and m stands for the sensitivity of the biosensor. Immunosensor based on a synergistic characteristic of organic–inorganic nanocomposite of rGO-hBNNS exhibited remarkably low LOD of 5.47 pg/mL offering a highly sensitive response for minute concentrations of CEA. Also, the performance of as fabricated rGO-hBNNS hybrid nanocomposite was found outstanding in comparison to various rGO-based immunosensors developed for biosensing application (Table 3).

Control, reproducibility, selectivity, stability, serum sample analysis

A series of studies were conducted to assess the practical applicability of the as-fabricated BSA/Anti-CEA/rGO-hBNNS/ITO nano-biosensor for real-time analysis. These studies aimed to evaluate the current reproducibility of immunoelectrodes fabricated under similar conditions, distinguish false responses, determine selectivity among various interferents present in real samples, assess stability over time, and analyze responses to spiked concentrations of CEA antigens in real samples.

Table 3 Comparison of sensing characteristics of rGO-hBNNS hybrid nanocomposite-based nano-biosensor with existing rGO and hBN-based immunosensors

Sensing matrix	Detection technique	Linear range (ng/mL)	Limit of detection	Sensitivity	Stability	Ref
APTES-nYZR	DPV	0.01–50	7.2 pg/mL	0.2 mA·ng^{-1}·mL·mm^{-2}	56 days	[40]
BN-nanosheets/GCE	EIS	0.1–500	0.017 ng/mL	17 pg/mL	60 days	[41]
PtNP(PPy-PPa)-RGO	EIS	10–1000	4 ng/mL	107.08 Ω·cm^2·$decade^{-1}$	30 days	[42]
rGO-PEI-Ag-Nf	DPV	0.5–500	0.38 ng/mL	25.29 µA·(log ng/mL)$^{-1}$	28 days	[43]
AuNP/PEI/rGO	CV, DPV	1–50	0.219 ng/mL	0.860 µA·mL·ng^{-1}·cm^{-2}	7 days	[44]
AuNP-rGO	DPV	0.0005–4	72.73 ng/mL	7.31 mA·mL·pg^{-1}	3 months	[45]
PS/rGO-MNP-PDA ENF	CV, DPV, EIS	0.5–60	0.33 ng/mL	0.354 µA·mL·ng^{-1}	-	[46]
L-methionine/ZrO_2-RGO	CV, DPV	1–20	0.079 ng/mL	4.8 µA·mL·ng^{-1}·cm^{-2}	9 weeks	[47]
Pd_{nano}/rGO/PEDOT	EIS, CV	0.2–20	0.2 ng/mL	0.375 mA·mL·ng^{-1}	8 weeks	[48]
nHfO_2-RGO	DPV	0–30	0.16 ng/mL	18.24 µA·mL·ng^{-1}	27 days	[49]
ZrO_2-RGO	CV, DPV	2–22	0.122 ng/mL	0.756 µA·mL·ng^{-1}	8 weeks	[39]
h-BNNS/ITO	CV, DPV	0–10	22.5 pg/mL	24.84 µA·mL·ng^{-1}·cm^{-2}	5 weeks	[21]
rGO-hBNNS/ITO	DPV, CV	0–15	5.47 pg/mL	9. 65 µA·mL·ng^{-1}·cm^{-2}	8 weeks	Present work

CV cyclic voltammetry, *DPV* differential pulse voltammetry, *EIS* electrochemical impedance spectroscopy, *nYZR* yttria-doped zirconia-reduced graphene oxide nanocomposite, *PS* polystyrene, *MNP* magnetic nanoparticle (Fe_3O_4), *PDA* polydopamine, *ENF* electrospun nanofibers

False responses, where the rGO-hBNNS/ITO electrode exhibits a response similar to that of the BSA/Anti-CEA/rGO-hBNNS/ITO immunoelectrode, were investigated. A control experiment was conducted to measure the current response upon immobilizing three different concentrations (0.1, 8, 15 ng/mL) of CEA onto three bare rGO-hBNNS/ITO electrodes, as depicted in Fig. 8a. The average current obtained from the control experiment was 0.177 mA, indicating no change in the current of the rGO-hBNNS/ITO electrode upon CEA immobilization. This result aligns well with the DPV current obtained in the electrode study (Table 2). The observed RSD of 0.08% in the control

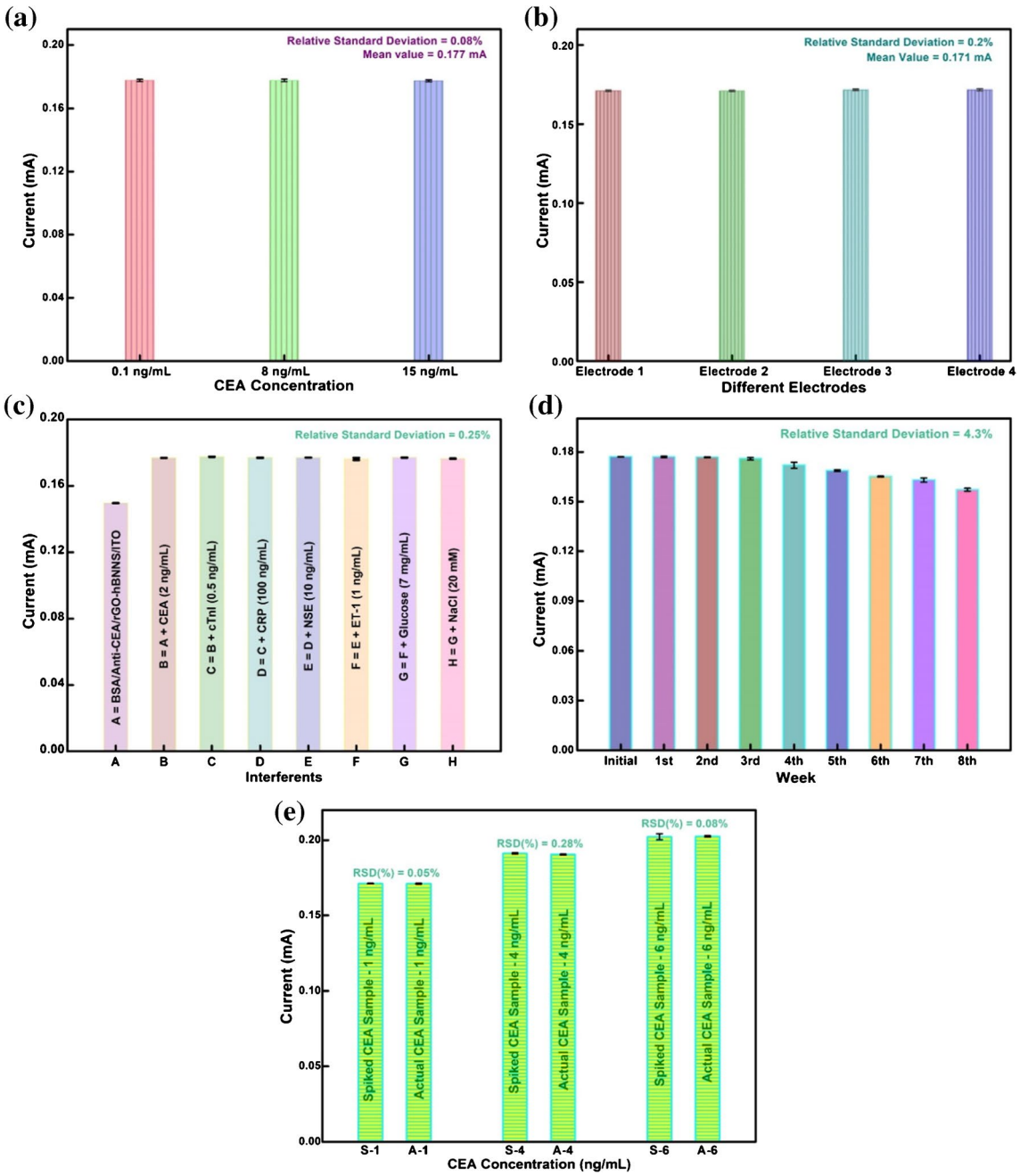


Fig. 8 Bar graph representing **a** control study of rGO-hBNNS/ITO electrode by addition of different CEA concentrations, **b** assessment of current reproducibility in as-fabricated immunoelectrode, **c** selectivity study conducted on successive modification of different interferent in electrode depicted as A = BSA/Anti-CEA/rGO-hBNNS/ITO, B = A + CEA(2 ng/mL), C = B + cTnI (0.5 ng/mL), D = C + CRP (100 ng/mL), E = D + NSE(10 ng/mL), F = E + ET-1 (1 ng/mL), G = F + Glucose (7 ng/mL), and H = G + Nacl (20 mM), **d** stability study performed weekly on BSA/Anti-CEA/rGO-hBNNS/ITO immunoelectrode, and **e** monitoring of spiked serum sample study utilizing by DPV technique

experiment emphasizes the consistent performance and reliability of the rGO-hBNNS/ITO electrode for protein biomarker detection, highlighting its suitability for accurate and precise measurements in this application.

A reproducibility study was conducted to assess the consistency of the electrochemical response in the as-fabricated immunosensor. DPV responses were monitored on four independent BSA/Anti-CEA/rGO-hBNNS/ITO immunoelectrodes after immobilizing 1 ng/mL CEA concentration onto each (Fig. 8b). The as-fabricated immunoelectrode exhibited minimal variation in DPV measurements, with an average value of 0.171 mA. Following the electrochemical investigation, the relative standard deviation (RSD) among the DPV currents measured from the four immunoelectrodes was determined to be 0.2% (< 5%), indicating outstanding reproducibility of the as-fabricated immunoelectrode. This suggests that the immunoelectrode consistently produces reliable results across multiple measurements, validating its suitability for consistent and accurate analysis in this context.

Human serum samples serve as vital clinical specimens for detecting various diseases characterized by diverse biological analytes. Therefore, assessing the specificity of the rGO-hBNNS-based immunosensor for carcinoembryonic antigen (CEA) amidst various interferents, such as cardiac troponin I (cTnI), C-reactive protein (CRP), neuron-specific enolase (NSE), endothelin-1 (ET-1), glucose, and sodium chloride (NaCl), is essential for its clinical utility. The presence of these analytes in serum samples may potentially impede the specific binding between CEA and the BSA/Anti-CEA/rGO-hBNNS/ITO immunoelectrode. Hence, investigating the selectivity of the rGO-hBNNS-based nano-biosensor was crucial, and this was accomplished by assessing the integrated effect upon successive addition of interfering analytes, as depicted in Fig. 8c via the DPV technique. Table S3, presented in the Electronic Supplementary Information, illustrates the outstanding selectivity of the BSA/Anti-CEA/rGO-hBNNS/ITO immunoelectrode towards CEA. The current response of the CEA/BSA/Anti-CEA/rGO-hBNNS/ITO immuno-interface remained unchanged even after the successive addition of interferents. The RSD of 0.25% highlights a highly selective response of the as-fabricated immunosensor. This implies that the immunosensor accurately identifies the target analyte (CEA) while remaining unaffected by the presence of interfering substances, showcasing its outstanding selectivity and reliability in practical applications.

The shelf life study was conducted using the DPV technique to assess the clinical storage stability of the CEA-immobilized BSA/Anti-CEA/rGO-hBNNS/ITO immunoelectrode (Fig. 8d). Testing was performed weekly to evaluate its suitability for real-time applications. During the interim period, the nano-biosensor was stored at 4 °C until subsequent examination. At the conclusion of the eighth week, the DPV current obtained was 88.74% of the initial DPV response, demonstrating a low relative standard deviation (RSD) of 4.3%, which is deemed acceptable (< 5%). These results indicate the robust storage stability of the fabricated immunosensor.

The human serum samples undergo clinical testing that validates their authenticity as real samples. Thus, quantitative determination of the spiked concentration of CEA in human serum samples that are known to exhibit various biological analytes examines the real-time applicability of as-fabricated immunoelectrode. In this study, the DPV response associated with the actual CEA concentration (A) was compared with the CEA concentration spiked into human serum samples (S) of identical concentration, which were then monitored (Fig. 8e).

A remarkable agreement in the average peak current of actual and spiked serum CEA concentrations was observed as shown in Table 4. The observed average peak currents associated with standard CEA concentrations of 1, 4, and 6 ng/mL were $0.171 \pm (5.5 \times 10^{-4})$ mA, $0.190 \pm (5.8 \times 10^{-4})$ mA, and $0.202 \pm (7.3 \times 10^{-4})$ mA, respectively. These values demonstrate a remarkable agreement with corresponding spiked CEA concentrations in human serum sample, where the average peak currents were found to be $0.171 \pm (3.7 \times 10^{-4})$ mA, $0.191 \pm (6.7 \times 10^{-4})$ mA, and $0.202 \pm (3.4 \times 10^{-3})$ mA, respectively. The RSDs for CEA concentrations of 1, 4, and 6 ng/mL were exceptionally low at 0.05%, 0.28%, and 0.08%, respectively. This indicates a high level of precision and consistency in the DPV current measurements obtained for these concentrations. Such low RSD values suggest minimal variability within the data sets, highlighting the reliability and accuracy of the BSA/Anti-CEA/rGO-hBNNS/ITO immunoelectrode towards detection of CEA in clinical serum samples of the patients as well.

Table 4 Recovery current and relative standard deviation (RSD, %) associated with detection of spiked concentration of CEA in human serum samples in comparison to actual CEA concentration

CEA concentration (ng/mL)	Average current for standard CEA sample (A, in mA)	Average current for spiked CEA sample (S, in mA)	RSD (%)	Recovery %
1	$0.171 \pm (5.5 \times 10^{-4})$	$0.171 \pm (3.7 \times 10^{-4})$	0.05	100.07
4	$0.190 \pm (5.8 \times 10^{-4})$	$0.191 \pm (6.7 \times 10^{-4})$	0.28	100.39
6	$0.202 \pm (7.3 \times 10^{-4})$	$0.202 \pm (3.4 \times 10^{-3})$	0.08	99.88

Conclusion

In summary, our present work harnesses the synergistic potential of organic–inorganic hybrid nanocomposites in fabricating nano-biosensing platform for ultrasensitive, efficient, cost-effective, and label-free detection of serum tumor marker associated with NSCLC. We observed an enhanced electrocatalytic performance (~29%) of hydrothermally synthesized rGO-hBNNS hybrid nanocomposite w.r.t. hBNNS owing to their unique crosslinking kinetics at interface. Combining amine-terminated hBNNS with carboxyl-functionalized rGO nanosheets created an environment abundant in surface active sites, facilitating the efficient immobilization of Anti-CEA antibodies. This enhancement significantly boosted the loading capacity of the nano-biosensor, resulting in exceptional sensitivity. Notably, this heightened sensitivity enabled the detection of CEA with an impressive low limit of 5.47 pg/mL. Furthermore, as-fabricated BSA/Anti-CEA/rGO-hBNNS/ITO immunoelectrode exhibited excellent storage stability for 8 weeks and was found to be highly selective in the presence of interfering markers present in serum. Their clinical applicability was unmatched owing to excellent recovery rates achieved for spiked serum sample analysis. This nano-biosensor depicted good analytical and clinical accuracy for monitoring potential biological analytes for real-time diagnostic applications.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00216-024-05379-6.

Acknowledgements We are greatly thankful to Prof. Prateek Sharma, Vice Chancellor, DTU, Delhi, India, for arranging the research facilities.

Author contribution Kanika Sharma: Conceptualization; methodology; data curation; investigation; validation; formal analysis; writing, original draft; writing, review and editing; visualization. Nitin K. Puri: Resources, supervision, funding acquisition, project administration, validation, writing—review and editing. Bharti Singh: Supervision, resources, validation.

Funding We gratefully acknowledged SERB, DST, India (EMR/2016/007479) for supporting this research.

Declarations

Ethics approval The human blood serum sample was obtained from All India Institute of Medical Sciences (AIIMS), New Delhi, with approval of the study.

Conflict of interest The authors declare no competing interests.

References

1. Alduais Y, Zhang H, Fan F, Chen J, Chen B. Non-small cell lung cancer (NSCLC): a review of risk factors, diagnosis, and treatment. Medicine. 2023;102(8):e32899-e.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: Cancer J Clin. 2021;71(3):209–49.
3. Frick C, Rumgay H, Vignat J, Ginsburg O, Nolte E, Bray F, et al. Quantitative estimates of preventable and treatable deaths from 36 cancers worldwide: a population-based study. Lancet Glob Health. 2023;11(11):e1700–12.
4. Jiang C, Zhao M, Hou S, Hu X, Huang J, Wang H, et al. The indicative value of serum tumor markers for metastasis and stage of non-small cell lung cancer. Cancers. 2022;14(20):5064.
5. Chinnappan R, Mir TA, Alsalameh S, Makhzoum T, Alzhrani A, Alnajjar K, et al. Emerging biosensing methods to monitor lung cancer biomarkers in biological samples: a comprehensive review. Cancers. 2023;15(13):3414.
6. Lee DS, Kim SJ, Kang JH, Hong SH, Jeon EK, Kim YK, et al. Serum carcinoembryonic antigen levels and the risk of whole-body metastatic potential in advanced non-small cell lung cancer. J Cancer. 2014;5(8):663.
7. Sanko V, Kuralay F. Label-free electrochemical biosensor platforms for cancer diagnosis: recent achievements and challenges. Biosensors. 2023;13(3):333.
8. Reddy KK, Bandal H, Satyanarayana M, Goud KY, Gobi KV, Jayaramudu T, et al. Recent trends in electrochemical sensors for vital biomedical markers using hybrid nanostructured materials. Adv Sci. 2020;7(13):1902980.
9. Baig N, Kammakakam I, Falath W. Nanomaterials: a review of synthesis methods, properties, recent progress, and challenges. Mater Adv. 2021;2(6):1821–71.
10. Batool A, Sherazi TA, Naqvi SAR. Organic–inorganic nanohybrid-based electrochemical biosensors. Hybrid Nanomaterials: Biomedical, Environmental and Energy Applications: Springer; 2022. p. 151–73.
11. Yang G, Zhu C, Du D, Zhu J, Lin Y. Graphene-like two-dimensional layered nanomaterials: applications in biosensors and nanomedicine. Nanoscale. 2015;7(34):14217–31.
12. Arora K, Srivastava S, Solanki PR, Puri NK. Electrochemical hydrogen gas sensing employing palladium oxide/reduced graphene oxide (PdO-rGO) nanocomposites. IEEE Sens J. 2019;19(18):8262–71.
13. Sandil D, Srivastava S, Malhotra B, Sharma S, Puri NK. Biofunctionalized tungsten trioxide-reduced graphene oxide nanocomposites for sensitive electrochemical immunosensing of cardiac biomarker. J Alloys Compd. 2018;763:102–10.
14. Sharma K, Puri NK. Enhanced electrochemical performance of hydrothermally exfoliated hexagonal boron nitride nanosheets for applications in electrochemistry. J Electrochem Soc. 2021;168(5):056512.
15. Maghraby YR, El-Shabasy RM, Ibrahim AH, Azzazy HME-S. Enzyme immobilization technologies and industrial applications. ACS Omega. 2023;8(6):5184–96.
16. Schroeder B, Le Xuan H, Völzke JL, Weller MG. Preactivation crosslinking—an efficient method for the oriented immobilization of antibodies. Methods Protoc. 2019;2(2):35.
17. Akouros A, Koutroumanis N, Manikas AC, Paterakis G, Carbone MGP, Anagnostopoulos G, et al. Highly stretchable strain sensors based on Marangoni self-assemblies of graphene and its hybrids with other 2D materials. Nanotechnology. 2023;34(29):295501.
18. Muratov DS, Vanyushin V, Koshlakova VA, Kolesnikov EA, Maksimkin AV, Stepashkin AA, et al. Improved mechanical and thermal properties of polypropylene filled with reduced graphene oxide (rGO) and hexagonal boron nitride (hBN) particles. J Alloys Compd. 2024;972:172882.
19. Mussa Y, Ahmed F, Arsalan M, Alsharaeh E. Two dimensional (2D) reduced graphene oxide (RGO)/hexagonal boron nitride

(h-BN) based nanocomposites as anodes for high temperature rechargeable lithium-ion batteries. Sci Rep. 2020;10(1):1882.
20. Byrappa K, Yoshimura M. Handbook of hydrothermal technology: William Andrew; Elsevier. 2012.
21. Sharma K, Puri NK, Singh B. An efficient electrochemical nanobiosensor based on hydrothermally engineered ultrathin nanostructures of hexagonal boron nitride nanosheets for label-free detection of carcinoembryonic antigen. Appl Nanosci. 2023;14(1):217–30.
22. Vivanco-Martinez F, Bragado R, Albar J, Juarez C, Ortiz-Masllorens F. Chemical modification of carboxyl groups in human Fcγ fragment: structural role and effect on the complement fixation. Mol Immunol. 1980;17(3):327–36.
23. Vashist SK. Comparison of 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide based strategies to crosslink antibodies on amine-functionalized platforms for immunodiagnostic applications. Diagnostics. 2012;2(3):23–33.
24. Welch NG, Scoble JA, Muir BW, Pigram PJ. Orientation and characterization of immobilized antibodies for improved immunoassays. Biointerphases. 2017;12(2).
25. Yang S, Yue W, Huang D, Chen C, Lin H, Yang X. A facile green strategy for rapid reduction of graphene oxide by metallic zinc. RSC Adv. 2012;2(23):8827–32.
26. Ferrari AC, Robertson J. Interpretation of Raman spectra of disordered and amorphous carbon. Phys Rev B. 2000;61(20):14095.
27. Pimenta M, Dresselhaus G, Dresselhaus MS, Cancado L, Jorio A, Saito R. Studying disorder in graphite-based systems by Raman spectroscopy. PCCP. 2007;9(11):1276–90.
28. Krishnamoorthy K, Veerapandian M, Yun K, Kim S-J. The chemical and structural analysis of graphene oxide with different degrees of oxidation. Carbon. 2013;53:38–49.
29. Purwaningsih H, Suari NMIP, Widiyastuti W, Setyawan H. Preparation of rGO/MnO2 composites through simultaneous graphene oxide reduction by electrophoretic deposition. ACS Omega. 2022;7(8):6760–7.
30. Yao Y-H, Li J, Zhang H, Tang H-L, Fang L, Niu G-D, et al. Facile synthesis of a covalently connected RGO–COF hybrid material by in situ reaction for enhanced visible-light induced photocatalytic H 2 evolution. J Mater Chem A. 2020;8(18):8949–56.
31. Hanifah MFR, Jaafar J, Othman M, Ismail A, Rahman MA, Yusof N, et al. Facile synthesis of highly favorable graphene oxide: effect of oxidation degree on the structural, morphological, thermal and electrochemical properties. Materialia. 2019;6:100344.
32. Debbarma R, Behura S, Nguyen P, Sreeprasad T, Berry V. Electrical transport and network percolation in graphene and boron nitride mixed-platelet structures. ACS Appl Mater Interfaces. 2016;8(13):8721–7.
33. Mathkar A, Tozier D, Cox P, Ong P, Galande C, Balakrishnan K, et al. Controlled, stepwise reduction and band gap manipulation of graphene oxide. J Phys Chem Lett. 2012;3(8):986–91.
34. Socrates G. Infrared and Raman characteristic group frequencies: tables and charts: John Wiley & Sons; 2004.
35. Martínez Cortizas A, López-Costas O. Linking structural and compositional changes in archaeological human bone collagen: An FTIR-ATR approach. Sci Rep. 2020;10(1):17888.
36. Suys O, Derenne A, Goormaghtigh E. ATR-FTIR biosensors for antibody detection and analysis. Int J Mol Sci. 2022;23(19):11895.
37. Sibai A, Elamri K, Barbier D, Jaffrezic-Renault N, Souteyrand E. Analysis of the polymer-antibody-antigen interaction in a capacitive immunosensor by FTIR difference spectroscopy. Sens Actuators, B Chem. 1996;31(1–2):125–30.
38. Elgrishi N, Rountree KJ, McCarthy BD, Rountree ES, Eisenhart TT, Dempsey JL. A practical beginner's guide to cyclic voltammetry. J Chem Educ. 2018;95(2):197–206.
39. Kumar S, Sharma JG, Maji S, Malhotra BD. Nanostructured zirconia decorated reduced graphene oxide based efficient biosensing platform for non-invasive oral cancer detection. Biosens Bioelectron. 2016;78:497–504.
40. Kumar S, Gupta N, Malhotra BD. Ultrasensitive biosensing platform based on yttria doped zirconia-reduced graphene oxide nanocomposite for detection of salivary oral cancer biomarker. Bioelectrochemistry. 2021;140:107799.
41. Öndeş B, Kilimci U, Uygun M, Uygun DA. Determination of carcinoembryonic antigen (CEA) by label-free electrochemical immunosensor using functionalized boron nitride nanosheets. Bioelectrochemistry. 2024;157:108676.
42. Puri N, Mishra SK, Niazi A, Srivastava AK, Rajesh. Physicochemical characteristics of reduced graphene oxide based Pt-nanoparticles-conducting polymer nanocomposite film for immunosensor applications. J Chem Technol Biotechnol. 2015;90(9):1699–706.
43. Wang Y, Ma D, Zhang G, Wang X, Zhou J, Chen Y, et al. An electrochemical immunosensor based on SPA and rGO-PEI-Ag-Nf for the detection of arsanilic acid. Molecules. 2021;27(1):172.
44. Liu X, Lin L-Y, Tseng F-Y, Tan Y-C, Li J, Feng L, et al. Label-free electrochemical immunosensor based on gold nanoparticle/polyethyleneimine/reduced graphene oxide nanocomposites for the ultrasensitive detection of cancer biomarker matrix metalloproteinase-1. Analyst. 2021;146(12):4066–79.
45. Verma S, Singh A, Shukla A, Kaswan J, Arora K, Ramirez-Vick J, et al. Anti-IL8/AuNPs-rGO/ITO as an immunosensing platform for noninvasive electrochemical detection of oral cancer. ACS Appl Mater Interfaces. 2017;9(33):27462–74.
46. Ketmen S, ErZeybekler S, Gelen SS, Odaci D. Graphene oxide-magnetic nanoparticles loaded polystyrene-polydopamine electrospun nanofibers based nanocomposites for immunosensing application of c-reactive protein. Biosensors. 2022;12(12):1175.
47. Gupta PK, Tiwari S, Khan ZH, Solanki PR. Amino acid functionalized ZrO 2 nanoparticles decorated reduced graphene oxide based immunosensor. J Mater Chem B. 2017;5(10):2019–33.
48. Gupta S, Jain U, Murti BT, Putri AD, Tiwari A, Chauhan N. Nanohybrid-based immunosensor prepared for Helicobacter pylori BabA antigen detection through immobilized antibody assembly with@ Pdnano/rGO/PEDOT sensing platform. Sci Rep. 2020;10(1):21217.
49. Kumar S, Kumar S, Augustine S, Yadav S, Yadav BK, Chauhan RP, et al. Effect of Brownian motion on reduced agglomeration of nanostructured metal oxide towards development of efficient cancer biosensor. Biosens Bioelectron. 2018;102:247–55.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Sustainable Water Resources Management (2024) 10:140
https://doi.org/10.1007/s40899-024-01123-1

ORIGINAL ARTICLE

Fe(III)-modified bamboo biochar for the removal of phosphate from synthetic and field stormwater runoff

Shivani Yadav[1] · Harsh Pipil[1] · A. K. Haritash[1] · Krishna R. Reddy[2]

Received: 28 February 2024 / Accepted: 3 June 2024
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2024

Abstract

Application of fertilizers over agricultural fields, domestic sewage, and livestock are major point and non-point sources of nutrients such as phosphate and nitrate in stormwater runoff. These nutrients, specifically phosphate, is the major pollutant triggering eutrophication in freshwater bodies, thus damaging both- the aquatic ecosystem and the environment. The current study investigates the removal of phosphate using biochar-based treatment units, which can be an efficient and cost-effective removal method. The biochar used in this study was produced by slow pyrolysis of bamboo and it was further chemically modified with 1%, 5%, and 10% $FeCl_3$ solutions. The iron modification transformed the biochar surface to be electro-positive. Characterization of both biochar and modified biochars was carried out using proximate analysis, elemental analysis, Fourier transform infrared spectroscopy, and scanning electron microscopy. Furthermore, a series of batch experiments were conducted to evaluate the phosphate and nitrate removal efficiencies from the synthetic stormwater runoff. The removal efficiencies were found to be 29%, 49%, and 61%, with an initial phosphate concentration of 5 mg/L, using 1%, 5% and 10% $FeCl_3$ treated biochar respectively, while the nitrate removal efficiencies (with an initial concentration of 50 mg/L) were found to be 12.4%, 14.6%, and 17.0% respectively. The phosphate removal kinetics by 10% $FeCl_3$ modified biochar indicated that the phosphate removal followed pseudo-second-order kinetics. Isotherm modelling indicated both Langmuir and Freundlich isotherms as best fit for phosphate adsorption using 10% $FeCl_3$ modified biochar thus representing mono-layer and multi-layer adsorption, respectively. Langmuir isotherm reported the maximum phosphate adsorption capacity of 11.57 mg/g by 10% $FeCl_3$ modified biochar. Additional experiments were performed using real urban stormwater runoff from various regions of Delhi, which revealed similar phosphate removal efficiencies. The treated stormwater runoff can be used for water capacity building and non-potable applications in water-scarce urban areas.

Highlights

- Pyrolysis of bamboo has 40% yield.
- Phosphates and nitrates in stormwater runoff can be effectively removed by biochar.
- Biochar requires its treatment with electro-negative chemical species for effective removal of nutrients.

Keywords Biochar · Stormwater runoff · Treatment · Eutrophication · Phosphate

Technical Paper Submitted to: *Sustainable Water Resources Management (SWAM)*

✉ Harsh Pipil
harsh_phd2k19@dtu.ac.in

Shivani Yadav
shivaniyadav_phd2k19@dtu.ac.in

A. K. Haritash
akharitash@dce.ac.in

Krishna R. Reddy
kreddy@uic.edu

[1] Department of Environmental Engineering, Delhi Technological University, Shahbad Daulatpur, Delhi 110042, India

[2] Department of Civil, Materials, and Environmental Engineering, University of Illinois, Chicago, IL 60607, USA

Published online: 17 June 2024

Introduction

The scarcity of water is becoming a challenge throughout the world leading to the crisis of potable water (He et al. 2021). India being the seventh largest country in the world and the most populous country, is rapidly advancing in terms of agricultural growth to feed its citizens (Ministry of Agriculture & Farmers Welfare 2022). Agricultural advancement also leads to the exploited application of chemical fertilizers in agricultural fields. These fertilizers consist of nutrients such as phosphate (PO_4^{3-}) and nitrate (NO_3^-) which when mixed with stormwater runoff degrades the quality of the receiving waterbody. In addition, phosphate is a critical pollutant causing eutrophication of receiving freshwater bodies (Haritash et al. 2017). The runoff from agricultural fields is a non-point source of nutrient pollution and causes excess nutrient loading leading to eutrophication, harmful algal blooms, and fish kills in receiving water bodies such as lakes and ponds (Sinha et al. 2017; Smith et al. 2020).

In addition, during dry seasons when no rainfall is being received in urban areas, pollutants such as pet wastes and biogenic wastes such as leaf litter and grass clippings get deposited on roads, streets, house roofs, trees, etc. During excess rainfall, the nutrients from the unutilized fertilizers, excess urban waste, and overflowing and leaking sanitary sewers get mixed with the stormwater runoff (Yang and Lusk 2018). The runoff traverses its flow path depending on the hydrology and imperviousness of the surface in that area (Kleinman 2006).

Nutrients in stormwater runoff are usually treated using several techniques such as biological treatment, ion exchange, ultrafiltration, membrane separation, and chemical precipitation. Employing some of these chemical techniques removes the pollutants through redox and ion exchange reactions, however, results in the release of harmful chemical species (Romero-Gueiza et al. 2015). Wetlands are also used to remove the nutrients from the receiving water, however, this method requires a large area and higher retention time moreover, the technique is not viable and efficient during the winter season (Haritash et al. 2015; Nandakumar et al. 2019; Pipil et al. 2021, 2023; Jain et al. 2024). Adsorption, on the other hand, is an effective technique that can be employed as a low-cost-efficient method to remove nutrients (Huang et al. 2020).

Among various adsorbents available, biochar, a product obtained from pyrolysis or gasification of biomass in the absence of oxygen, can be an efficient low-cost adsorbent owing to several favourable properties such as high specific surface and internal porosity (Weber and Quicker 2018). The adsorption capacity of biochar depends on several factors like the functional group, the surface charge, and the specific surface area. Few studies were conducted in the past to explore the potential removal of nutrients from wastewater using biochar (Zhang et al. 2020; Reddy et al. 2014). The surface of biochar is negatively charged which makes it unsuitable for the effective removal of anionic chemical species such as phosphate and nitrate, and thus, requires treatment with cationic chemical species like Fe(III) to effectively remove these nutrients from runoff (Mukherjee et al. 2011). $FeCl_3$ can effectively remove phosphate from the aqueous medium. Fe-P coprecipitate shows high stability under high pH fluctuation ($3.0 < pH < 11.0$). Moreover, $FeCl_3$ acts as an alternate terminal electron acceptor oxidising the other chemical species present in aqueous medium (Li et al. 2019).

The characterisation of stormwater runoff is necessary to estimate the suspended solids, pollutants, and nutrient loading in the runoff so that the requisite and necessary treatment can be provided for it. In previously published studies, the stormwater runoff samples were collected from various land uses based on associated activity in Delhi, India, which includes commercial, industrial, institutional, residential, and road/highways (Pipil et al. 2022a). Study shows phosphate concentration ranges between 0.04 and 17.4 mg/L while nitrate concentrations range between 1.6 and 478.8 mg/L. The higher values of phosphates and nitrate can be attributed to the addition from organic matter, mulch, applied compost, etc. in lawns and parks, littering of food waste, and the excretory waste of stray animals and pets. Water-sensitive urban design (WSUD) such as rain gardens and wetlands can remove pollutants such as suspended solids, nutrients, and heavy metals when amended with an adsorbent such as biochar (Pipil et al. 2022b).

In the present study, biochar was produced by pyrolysis using bamboo and was subsequently impregnated with Fe(III). A detailed characterization of the biochar and Fe(III)-modified biochar was performed. A series of batch adsorption experiments were conducted using bamboo-biochar and Fe(III)-modified biochar to determine their phosphate removal kinetics and maximum adsorption capacity. Synthetically prepared stormwater consisting of phosphate and nitrate was subjected to treatment process using biochar. Furthermore real stormwater runoff consisting of phosphate and multiple co-existing chemicals was also treated using the same. The experimental results helped to assess the extent of phosphate removal and the mechanisms responsible for the removal.

Materials and methods

Preparation of biochar

Old refused bamboo was collected from the university campus which was cleaned thoroughly with Type-I water. The pieces were cut out of it to prepare biochar in bulk quantity to have homogeneous characteristics. The bamboo samples were then dried in a hot air oven for 24 h at 105 °C. The dried samples were then placed inside a muffle furnace at 500 °C for 30 min for pyrolysis in an air-tight chamber to ensure the presence of zero oxygen. The temperature was raised in the muffle at a rate of 16.67 °C/minute for 30 min to attain a final pyrolysis temperature of 500 °C. The biochar so formed was allowed to cool freely and naturally inside the muffle. The produced biochar was crushed and passed through IS 4.75 mm sieve but retained on the IS 2.36 mm sieve, which was the chosen grain size range for treatment. A solution of $FeCl_3$ of strength 1%, 5%, and 10% was prepared in which an equal quantity of prepared biochar was added. It was subjected to shaking for 24 h in a bench-top shaker at a RPM of 150 and the temperature was set at 27 °C. This biochar impregnated with Fe(III) was then washed thrice thoroughly with Type-I water. It was then dried in a hot air oven at 105°C for 24 h. In addition, one set of experiments was conducted with untreated biochar. The untreated biochar is indicated with B0%, while biochar having 1%, 5%, and 10% $FeCl_3$ was named as B1%, B5%, and B10%, respectively. The percentage yield and the cost associated with the preparation of biochar were also calculated.

Characteristics of biochar

Proximate analysis

To estimate the moisture, volatile, fixed carbon, and ash content of both untreated biochar and raw bamboo, proximate analysis was conducted following ASTM (ASTM E871-82, ASTM E1755-01, ASTM E872-82) standard methods using ELTRA - Thermogravimetric Analyzer (manufactured in Germany). The temperature for moisture, volatile, fixed carbon, and ash content was fixed at 105 °C, 915 °C, 750 °C, and 750 °C, respectively. The thermogravimetric analysis was conducted using 1100 ± 10 mg of samples in the crucibles.

Elemental analysis

Elemental analysis was performed for the determination of carbon, hydrogen, nitrogen, and sulphur (CHNS) using Eurovector - EuroEA3000 Series elemental analyser at 950 °C. The calibration was performed using L-Cystine as the reference standard material in tin capsules. CHNS analysis was conducted in duplicates using ≈ 2.5 mg samples in tin capsules.

Fourier transform infrared (FTIR) spectroscopy

A non-destructive method was used to identify the various functional groups present in biochar. The FTIR was conducted for untreated biochar and Fe(III) modified biochar before and after the experiments, to characterize the functional groups present. A Thermo Fisher Nicolet spectrometer enabled with a KBr beam splitter was used in the mid-infrared region. All the samples were ground in a mortar pestle which was cleaned thoroughly with an organic solvent. Sample pellets were prepared in the ratio of 1:10 with KBr. Spectra of KBr were run initially to cancel out its effect and its spectrum in the final results. The spectra were collected in the range of 400 cm^{-1} and 4000 cm^{-1} with a resolution of 2 cm^{-1}.

Scanning electron microscopy (SEM)

A non-destructive method for Scanning Electron Microscopy was performed using Zeiss EVO 18 special (Zeiss, Germany). The biochar samples were used to analyse their structure before and after the adsorption in the experiment. The samples were dried in a hot air oven for 24 h and then it was visualised on the copper-coated grid after gold coating between 750x and 1500x magnification.

Experimental setup

Three combinations of concentration for both phosphate (PO_4^{3-}-P) and nitrate (NO_3^--N) were selected while preparing the mixed synthetically prepared stormwater runoff. Phosphate and nitrate were present in the mixed synthetically prepared stormwater having the concentration of 2.0 mg/L PO_4^{3-}-P, 25.0 mg/L NO_3^--N; 5.0 mg/L PO_4^{3-}-P, 50.0 mg/L NO_3^--N; and 10.0 mg/L PO_4^{3-}-P, 75.0 mg/L NO_3^--N. Initially, a dose of 1.0 g/L of biochar was introduced for the treatment which was later increased to 5.0 g/L. The experiments were conducted in triplicates in a conical flask with 200 mL of synthetically prepared stormwater runoff and the respective doses of biochar. The sample with biochar was subjected to shaking at 150 RPM at 27 °C. An aliquot of 1.0 mL was extracted at every 1-hour time interval. pH, electrical conductivity, and total dissolved solids (TDS) were noted using a Labman make (LMMP 30 model) bench-top multiparameter before and after the experiment. The residual concentrations of phosphate and nitrate were analysed using single beam spectrophotometer (Labtronics

make LT-290 Model) at $\lambda_{max} = 690$ nm for phosphate, and at $\lambda_{max} = 220$ nm for nitrate.

Removal of nutrient

The removal efficiency of nutrients from synthetically prepared and real stormwater runoff were calculated using the Eq. (1) given below:

$$\%Removal\ efficiency = \left[\frac{C_i - C_e}{C_i}\right] \times 100 \quad (1)$$

Where C_i is the initial concentration of nutrients in stormwater runoff in mg/L; C_e is the concentration of nutrients at equilibrium after treatment in mg/L. The removal efficiency was calculated based on the extraction of samples at an interval of 1 h which was extracted for 4 h.

Results and discussion

Characterisation of biochar

Proximate analysis

Based on thermogravimetric analysis, raw bamboo has more moisture and volatile content while its biochar has more carbon and ash content as compared to raw bamboo (Table 1). Raw bamboo has moisture, fixed carbon, volatile, and ash content as 7.1%, 16.8%, 76.1%, and 5.1%, respectively. On the other hand, bamboo biochar has moisture, fixed carbon, volatile, and ash content of order 3.5%, 59.6%, 17.4%, and 16.8%, respectively. During pyrolysis, the temperature was raised from room temperature to a peak 500 °C, resulting in the loss of moisture and water of hydration in the form of vapours, while the sample could have under decomposition which released complex organic compounds including CO_2, CO, H_2, CH_4 and other hydrocarbons in the form of syngas. This might have reduced the volatile content in biochar significantly. For the same reason of loss of moisture and water of hydration, the moisture content also reduced to half. The substantial increase in the fixed carbon content of bamboo biochar compared to raw bamboo is a result of the loss of other volatile organics during the pyrolysis (Sun et al. 2014).

Elemental analysis

Based on elemental analysis, as presented in Table 1, it was observed that pyrolysis has significantly changed the elemental composition of biochar. The carbon content of bamboo biochar increased to 73.2% from the 43.0% as compared to raw bamboo. However, the hydrogen and oxygen content in the biochar was reduced from 7.3 to 3.9% and 48.8–22.4%, respectively (Table 1). Nitrogen and sulphur content remained < 1.0% for both raw bamboo and bamboo biochar. O/C and H/C ratios reduced from 1.13 to 0.31 and 0.17 to 0.05, respectively, on the conversion of bamboo biomass to biochar. The lower ratio of O/C and H/C impart greater stability and degree of aromaticity within the biochar (Kumar et al. 2013).

Fourier transform infrared (FTIR) spectroscopy

As a result of pyrolysis, the increase in temperature changes the functional groups in the biochar. Non-destructive FTIR spectroscopy shows the peak intensity of the various organic groups commonly observed in biochar (Table 2). Between 3550 cm^{-1} to 3200 cm^{-1}, organic O-H stretching is a result of contribution from water retained in the biochar. C-H stretching between 3000 cm^{-1} to 2840 cm^{-1} corresponds to the presence of an alkane group. The peak between 2000 cm^{-1} to 1650 cm^{-1} is C = O stretching which consists of aromatic compounds, ketones, and amides for all the types of biochar. FTIR peak from 1400 cm^{-1} to 1000 cm^{-1} corresponds to C-H bending and C-O stretching that corresponds to the aldehyde and alcohol functional groups, respectively. 1000 cm^{-1} to 650 cm^{-1} corresponds to C = C stretching and C-H bending that relates to the alkene functional group (Fig. 1).

Scanning electron microscopy (SEM)

SEM images of biochar samples show the surface morphology of the biochar before and after treatment with $FeCl_3$. In both the biochar, complex surface morphology can be seen along with macro to micropores present over the surface

Table 1 Elemental characteristics of bamboo and its biochar

Particular	Proximate Analysis				Elemental Analysis						
	Moisture (%)	Fixed Carbon (%)	Volatile (%)	Ash (%)	Carbon (%)	Hydrogen (%)	Nitrogen (%)	Sulphur (%)	Oxygen* (%)	O/C ratio	H/C ratio
Raw bamboo	7.1	16.8	76.1	5.1	43.0	7.3	0.9	0.01	48.8	1.13	0.17
Bamboo biochar	3.5	59.6	17.4	16.8	73.2	3.9	0.5	0.001	22.4	0.31	0.05

*Calculated on the basis of weight difference of elements during analysis assuming it consists of these elements only

Table 2 Summary of FTIR functional group present in biochar samples

Functional group	B10% used biochar	B10% unused biochar	B0% used biochar	B0% unused biochar
	FTIR peak position Intensity	FTIR peak position Intensity	FTIR peak position Intensity	FTIR peak position Intensity
O-H stretch (alcohol)	3445 cm^{-1} Intensity: Strong, broad	3413 cm^{-1} Intensity: Strong, broad	3403 cm^{-1} Intensity: Strong, broad	3439 cm^{-1} Intensity: Strong, broad
C-H stretch (alkane)	2975 cm^{-1} Intensity: Medium	2917 cm^{-1} Intensity: Medium	2971 cm^{-1} Intensity: Medium	2920 cm^{-1} Intensity: Medium
C = O stretch (conjugated aldehyde)	1694 cm^{-1} Intensity: Strong	1620 cm^{-1} Intensity: Strong	1700 cm^{-1} Intensity: Strong	1692 cm^{-1} Intensity: Strong
C = C stretch (unsaturated ketone)	1594 cm^{-1} Intensity: Strong	Peak not observed	1588 cm^{-1} Intensity: Strong	Peak not observed
C-H bending (alkane)	Peak not observed	1437 cm^{-1} Intensity: Medium	Peak not observed	Peak not observed
C-H bending (aldehyde)	1385 cm^{-1} Intensity: Medium	1386 cm^{-1} Intensity: Medium	1386 cm^{-1} Intensity: Medium	1383 cm^{-1} Intensity: Medium
C-O stretch (tertiary alcohol)	1151 cm^{-1} Intensity: Strong	Peak not observed	1163 cm^{-1} Intensity: Strong	1116 cm^{-1} Intensity: Strong
C-O stretch (secondary alcohol)	1109 cm^{-1} Intensity: Strong	Peak not observed	1127 cm^{-1} Intensity: Strong	Peak not observed
C = C stretch (alkene)	961 cm^{-1} Intensity: Strong	Peak not observed	952 cm^{-1} Intensity: Strong	Peak not observed
C-H bending	882 cm^{-1} Intensity: Strong	812 cm^{-1} Intensity: Strong	817 cm^{-1} Intensity: Strong	806 cm^{-1} Intensity: Strong

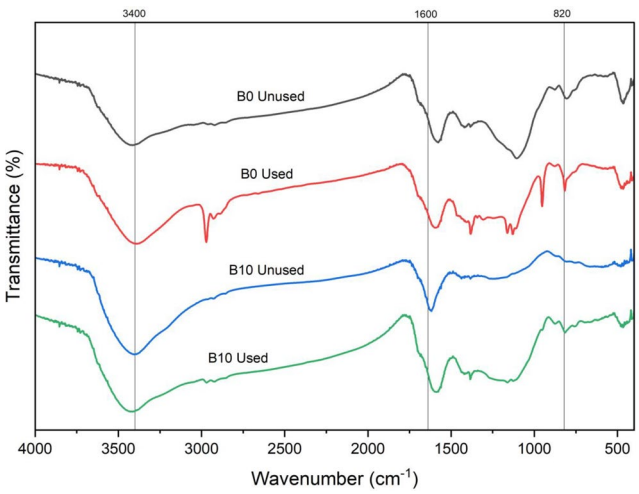


Fig. 1 FTIR spectrum of treated and untreated biochar before and after use

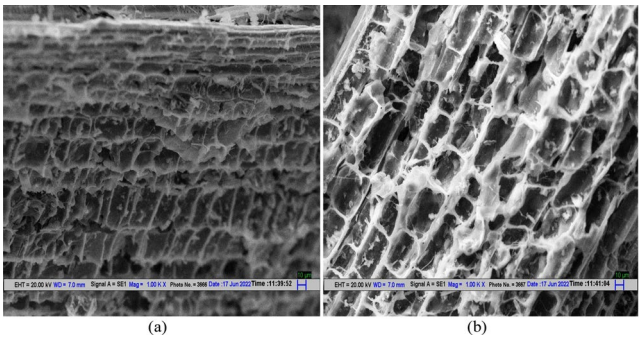


Fig. 2 SEM image of the biochar (a) before and (b) after impregnated with $FeCl_3$

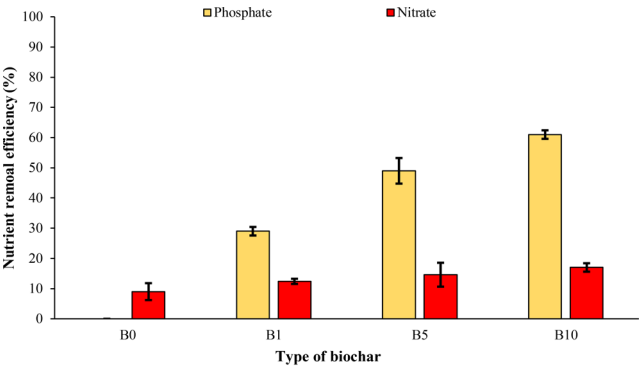


Fig. 3 Phosphate and nitrate removal efficiency of biochar from synthetically prepared stormwater runoff having PO_4^{3-} = 5 mg/L and NO_3^- =50 mg/L strength and biochar dose of 5.0 g/L

which are produced as a result of a vascular bundle of raw bamboo biomass (Hernandez-Mena et al. 2014). However, in biochar treated with $FeCl_3$, agglomeration of this electropositive chemical can be seen (Fig. 2).

Removal of nutrients

Removal from synthetic stormwater runoff

For the phosphate strength of 5 mg/L, different types of biochars have shown different removal efficiencies. B0% biochar without any treatment was not able to remove the phosphate from synthetically prepared stormwater runoff in a period of 4 h. This could be attributed to the non-availability of adsorption sites on the surface of untreated and negatively charged biochar. Negatively charged sites cannot remove anionic species due electrostatic force of repulsion. However, biochar which was pre-treated with $FeCl_3$ was able to remove phosphate. B1%, B5%, and B10% biochars removed 29%, 49%, and 61%, respectively, from synthetically prepared stormwater runoff when initial phosphate strength was 5.0 mg/L (Fig. 3). B10% treated biochar removed phosphate almost twice the B1% treated biochar. This could be attributed to the availability of sites having

Fe(III) present on the porous surface of biochar that can easily remove phosphate from the sample. The increase in phosphate removal efficiency is almost 60 times for B10% treated biochar to that of the untreated B0% biochar in the present study. B10% biochar was further used to study its nutrient removal efficiency for different phosphate strengths of 2 mg/L, 5 mg/L, and 10 mg/L, and the results are presented in Fig. 4. It was observed that biochar can remove almost 63% of phosphate for the initial phosphate strength of 2 mg/L at a dose of 5.0 g/L absorbent and a contact period of 4 h. At the same time, the removal efficiency of phosphates reduced with the increase in the initial strength of phosphate (Fig. 4). For the same dose and contact period, only 59% and 49% of phosphate were removed from synthetically prepared stormwater runoff of strength 5 mg/L and 10 mg/L, respectively. On the other hand, for the initial strength of nitrate as 50 mg/L, biochar without any given treatment was able to remove approximately 9%. Increasing Fe concentration in biochar did not show significant removal efficiency towards nitrate. The removal efficiency of nitrate using B1%, B5%, and B10% was of order 12.4%, 14.6%, and 17.0%, respectively. The lower adsorption of nitrate compared to phosphate can be due to the preferential adsorption of phosphate over nitrate by Fe(III) modified biochar.

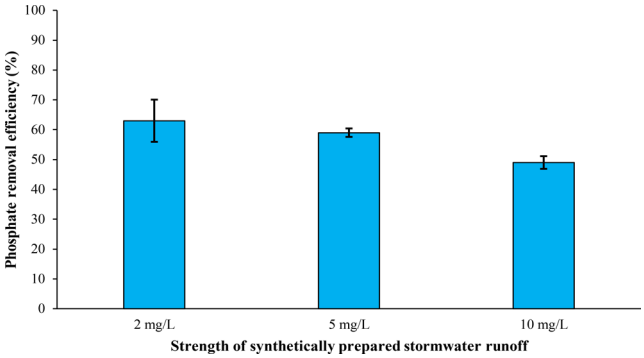


Fig. 4 Phosphate removal efficiency using B10% biochar from synthetically prepared stormwater runoff of varying strength for biochar dose of 5.0 g/L

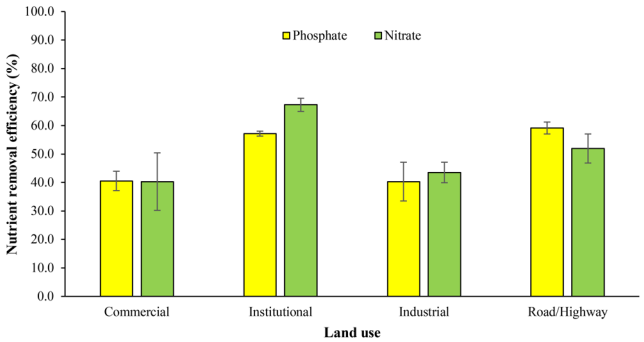


Fig. 5 Phosphate and nitrate removal efficiency (%) from stormwater runoff originating from different land use areas using B10% biochar

Removal from field stormwater runoff

Based on the observations obtained for nutrient removal from synthetically prepared stormwater, real (field) stormwater runoff was treated using B10% biochar, to verify its adsorption capacity in chemically complex real stormwater runoff. The phosphate removal efficiency of biochar for stormwater runoff collected from commercial, institutional, industrial, and road/highway areas of the Delhi region was of order 40.5%, 57.1%, 40.3%, and 59.1%, respectively (Fig. 5). It shows that the removal efficiency observed is relatively similar to that of the synthetically prepared stormwater runoff experiments. Also, up to 67.3% of nitrate was removed from institutional area runoff. The removal efficiency depends upon the initial concentrations of the pollutants. The equilibrium of the reaction shifts towards another side that reverses and releases pollutants from the adsorbed surface of biochar. The shift of equilibrium depends upon the release of pollutants from the solid phase (biochar) to an aqueous phase.

Adsorption capacity

The adsorption capacity q_e (mg/g) of the biochar was calculated using the Eq. (2)

$$q_e = \left[\frac{C_i - C_e}{W}\right] \times V \tag{2}$$

Where q_e is the adsorption capacity of biochar in mg/g; C_i is the initial concentration of nutrients in stormwater runoff in mg/L; C_e is the concentration of nutrients at equilibrium after treatment in stormwater runoff in mg/L; V is the volume of sample used for analysis in litre; and W is the quantity of absorbent in g. The adsorption capacity of biochar for a dose of 5.0 g/L was calculated in batch experiments for varying strength of phosphate in synthetic and real stormwater runoff which varied from 2.0 mg/L to 10 mg/L, and 0.4 mg/L to 3.2 mg/L, respectively. The experiment was conducted at 7.0 pH for synthetically prepared stormwater runoff at 27 °C temperature while the natural pH of real stormwater runoff is considered for this study. The dose of biochar was kept fixed at 5.0 g/L for Fe(III) treated B10% biochar for experimentation. The adsorption capacity of biochar increased with the increase in the concentration of phosphate which may be attributed to the availability of vacant sites for its adsorption which is represented by a strong correlation value ($R^2 > 0.99$) as shown in Fig. 6.

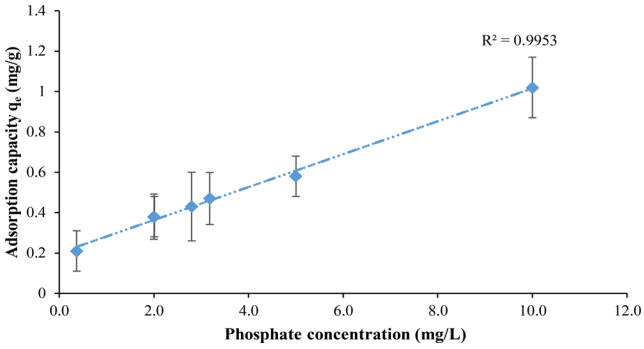


Fig. 6 Adsorption capacity of biochar as calculated during removal of phosphate from synthetic and real stormwater runoff tested in a batch experiment

Mechanism of removal

Removal of chemical species by biochar depends upon the presence of adsorption sites over the surface which is correlated with the specific surface area of the biochar produced. The increase in temperature for biochar production by pyrolysis, the specific surface area of biochar increases that enhances the removal of phosphates (Zeng et al. 2013). However, it was reported in earlier studies that biochar has an electro-negative charge over its surface, and it cannot remove anions (PO_4^{3-} and NO_3^-) by adsorption (Wang et al. 2015; Zeng et al. 2013). The pyrolysis of biomass produces biochar having functional groups such as -OH and -COOH as shown by the FTIR spectrum peak at 3400 cm^{-1} and 1600 cm^{-1}, respectively (Fig. 1), which may result in a negative surface charge. The net negative surface charge may lead to poor electrostatic attraction of PO_4^{3-} and NO_3^- with biochar and thus, require treatment with electropositive chemical species like $FeCl_3$. The impregnation of $FeCl_3$ over the biochar surface has played a role in nutrient removal. The SEM images before and after the adsorption (Fig. 2) revealed that the surface was smooth before adsorption while, after adsorption, the surface was occupied with adsorbed particles. This change in the surface morphology of biochar suggests that nutrients were removed by the action of electrostatic attraction between positively charged Fe(III) ions and negatively charged phosphate ions. FTIR spectrum peak at 1000 cm^{-1} after adsorption suggests the interaction of Fe-OH with P ions (Zhang et al. 2009).

Kinetics of nutrient removal

To find the kinetics of nutrient removal, pseudo-first-order and pseudo-second-order models were applied. Pseudo-first-order kinetic model is given by Eq. (3)

$$\log(q_e - q_t) = logq_e - \frac{K_1 t}{2.303} \tag{3}$$

Where q_e is the adsorption capacity of biochar in mg/g; q_t is the adsorption capacity of biochar in mg/g at any time t; K_1 is pseudo-first order rate constant (min^{-1}).

Pseudo-second-order kinetic model is given by Eq. (4)

$$\frac{t}{q_e} = \frac{1}{K_2 q_e^2} + \frac{t}{q_e} \tag{4}$$

Where q_e is the adsorption capacity of biochar in mg/g; K_2 is pseudo-second-order rate constant (g mg^{-1} min^{-1}) and t is time in minutes. The plots of pseudo-first-order and pseudo-second-order models for the adsorption capacity of biochar are shown in Fig. 7a and b respectively. The various parameters obtained from these models are given in Table 3. Pseudo-first-order model fits relatively poor for the removal of phosphates in which the R^2 value varied from 0.867 to 0.945. Also, the adsorption capacity calculated at equilibrium ($q_{e(cal)}$) and adsorption capacity from the experiments ($qe_{(exp)}$) varied significantly. Thus, it shows that adsorption kinetics does not follow the pseudo-first-order kinetic model. On the other hand, the R^2 value varied between 0.999 and 0.996 for pseudo-second-order kinetics which shows that the plot fits very well. Also, adsorption capacity varied between 0.35 and 0.60 mg/g for phosphate. Thus, the adsorption of phosphate follows pseudo-second-order rather than pseudo-first-order kinetics as reported in another study (Zhang et al. 2012).

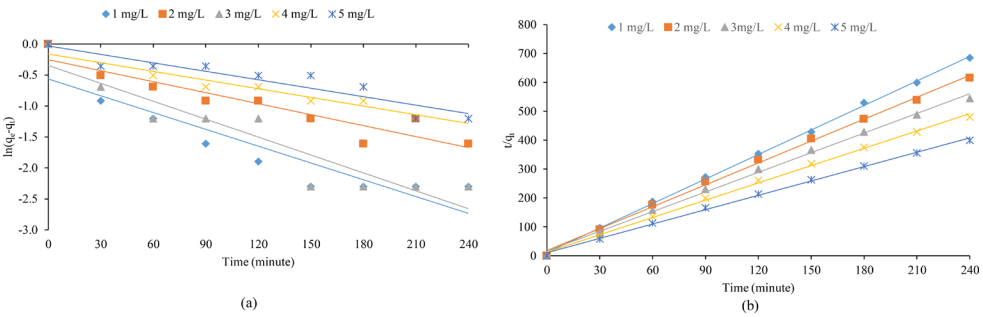


Fig. 7 (a) Pseudo first order kinetics, (b) Pseudo second order kinetics for removal of varying phosphates concentration using B10% biochar

Table 3 Comparison of parameters for pseudo-first order and pseudo-second order reaction kinetics

Order of reaction	Parameter	Unit	Concentration				
			1.0 mg/l	2.0 mg/l	3.0 mg/l	4.0 mg/l	5.0 mg/l
-	$q_{e(exp)}$	mg/l	0.31	0.38	0.45	0.51	0.58
First Order	K_1	min^{-1}	-0.009	-0.0059	-0.0096	-0.0047	-0.0045
	$q_{e(cal)}$	mg/g	0.57	0.78	0.71	0.85	0.97
	R^2	-	0.8431	0.8727	0.8724	0.9457	0.8676
Second order	K_2	g mg^{-1} min^{-1}	0.73	0.35	0.31	0.31	0.27
	$q_{e(cal)}$	mg/g	0.35	0.40	0.44	0.50	0.60
	R^2	-	0.9991	0.9973	0.9968	0.9976	0.9981

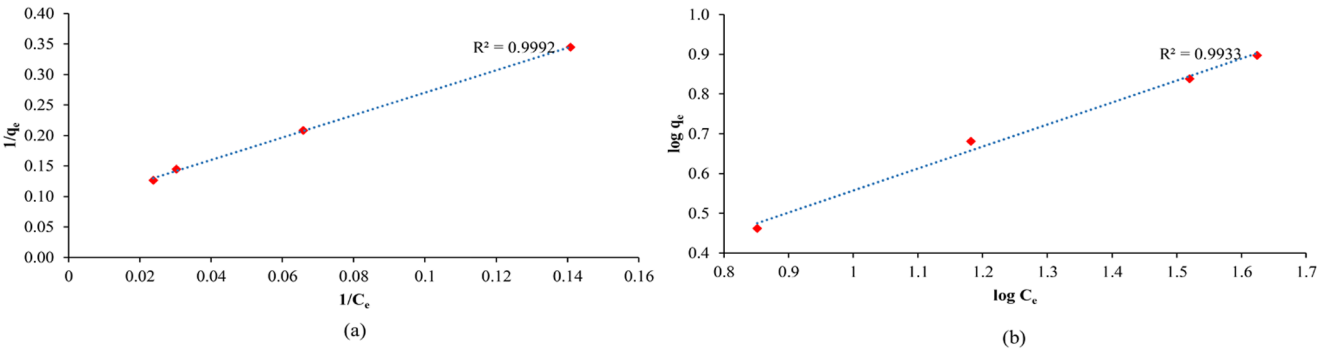


Fig. 8 (a) Langmuir isotherm; and (b) Freundlich isotherm fit for B10% biochar

Adsorption Isotherm

To find the adsorption isotherm, both Langmuir and Freundlich isotherm models were applied. Langmuir isotherm in its linear form is represented by Eq. (5)

$$\frac{1}{q_e} = \frac{1}{K_L q_{max}} \frac{1}{C_e} + \frac{1}{q_{max}} \tag{5}$$

Where q_e is the adsorption capacity of biochar in mg/g; q_{max} is the maximum adsorption capacity of biochar in mg/L; K_L is Langmuir's isotherm constant (L/mg).

The separation factor R_L is given by Eq. (6)

$$R_L = \frac{1}{1 + C_i \times K_L} \tag{6}$$

Where, R_L is Langmuir constant which indicates the adsorption possibility which may be either favourable ($0 < R_L < 1.0$), unfavourable ($R_L > 1.0$), linear ($R_L = 1.0$) or irreversible ($R_L = 0$).

Fruendlich isotherm in its linear form is represented by Eq. (7):

$$log q_e = log K_f + \frac{1}{n} log C_e \tag{7}$$

Where q_e is the adsorption capacity of biochar in mg/g; K_f is the Freundlich's constant $((mg.g^{-1})\ (L.mg^{-1})^{1/n})$; C_e is the concentration of nutrient at equilibrium after treatment in stormwater runoff in mg/L. In general, adsorption isotherm distinguishes the surface properties and affinity to adsorb the adsorbate. Isotherms corresponding to B10% biochar were studied to understand the same. The best curve was obtained by the Langmuir isotherm model with $R^2 = 0.9992$ as against the Freundlich isotherm model with $R^2 = 0.9933$ (Fig. 8). The maximum adsorption capacity q_{max} was obtained as 11.57 mg/g for phosphate (Table 4). The calculated value of Langmuir constant R_L was less than 1.0 which suggests that the adsorption possibility is favourable (Ayub et al. 2020). A brief comparison of studies done in the past for the removal of phosphate is summarized in Table 5.

Table 4 Langmuir and Freundlich adsorption isotherm for B10% biochar

Isotherm	Parameter	Unit	Values
Langmuir	q_{max}	(mg/g)	11.5741
	K_L	(L/mg)	0.0469
	R_L	-	0.3473
	R^2	-	0.9992
Freundlich	K_f	$(mg/g)\ (L/mg)^{1/n}$	1.0097
	1/n	-	0.5531
	R^2	-	0.9933

Yield of biochar

The yield of biochar produced during the experiment was calculated as the ratio weight of carbonaceous biochar

Table 5 Comparison of adsorption capacity of different materials in different studies for removal of phosphate

Adsorbent material	Adsorption capacity for PO_4^{3-} (mg/g)	Reference
Mg corn-cob biochar	56.12	Deng et al. 2021
Al corn-cob biochar	44.79	Deng et al. 2021
Pig manure biochar	9.51	Wang et al. 2015
lanthanum (La) oak biochar	46.37	Wang et al. 2016
Coal Gangue/biochar	3.19	Qiu and Duan 2019
Fly ash/biochar	3.07	Qiu and Duan 2019
Fe(III) Bamboo biochar	**11.57**	**Present study**

produced after pyrolysis to the weight of bamboo biomass before pyrolysis as given by the Eq. (8)

$$\%\ Biochar\ yield = \frac{weight\ of\ carbonaceous\ biochar\ produced\ after\ pyrolysis}{weight\ of\ bamboo\ biomass\ before\ pyrolysis} \times 100 \quad (8)$$

Using the formula mentioned above, approximately 40% biochar was produced from the initial bamboo biomass. The 60% of the initial mass loss can be attributed to the removal of moisture, the production of syngas, and the production of bio-oil during pyrolysis (Panwar and Pawar 2020). Biomass with higher water content will lead to a poor yield of biochar.

Cost of biochar production

Biochar was made through pyrolysis in a muffle furnace having a power rating of 2.0 KW. The muffle furnace was operated for an hour, initial 30 min for raising the temperature at a rate of 16.67 °C/minute, and the next 30 min for the actual process of pyrolysis at 500 °C. Thus, 2.0 KW muffle operated for 1 h implies that 2 units of electricity, i.e., 2.0 kWh was consumed in this process. The average price of electricity units in Delhi ₹6.0 per kWh. Thus, electricity cost accounts for ₹12.0 (Table 6). $FeCl_3$ was also used to impregnate the biochar in various configurations. The cost of $FeCl_3$

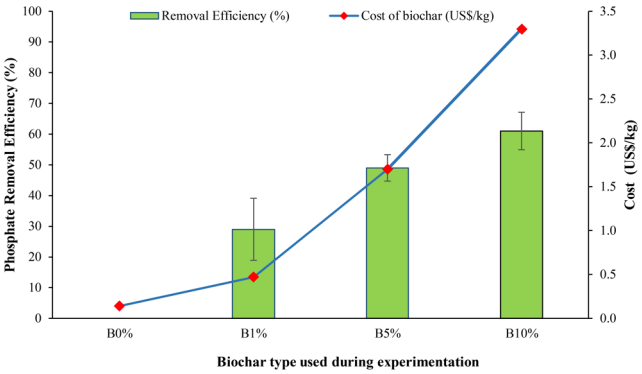


Fig. 9 Removal efficiency (%) and associated cost (US$) of different types of biochar (per kg) for phosphate removal from synthetically prepared stormwater runoff with initial PO_4^{3-}-P concentration is 5 mg/L

used during experimentation was ₹400 per 500 g packaging. Thus, the total cost of biochar includes the cost of electricity used during pyrolysis and the cost of chemicals consumed. Hence, the cost of biochar varies between ₹12.0 per kg for non-impregnated biochar to ₹244.0 per kg for 10% $FeCl_3$ impregnated biochar. Considering the currency exchange rate as US$1 equivalent to ₹83.17, the cost of biochar production becomes equivalent to US$ 0.14 per kg to US$ 2.93 per kg for the present study (Fig. 9).

Conclusions

The current study aimed to evaluate the enhanced phosphate and nitrate removal efficiency by Fe(III) modified biochar from synthetically prepared stormwater runoff containing phosphate and nitrate, and from actual stormwater runoff collected from the field. The biochar used in this study was produced by slow pyrolysis of bamboo biomass at a yield rate of 40%. The bamboo-biochar was then treated with 1%, 5%, and 10% $FeCl_3$ solutions to induce surface electro-positivity. Biochar and modified biochars were then characterized based on proximate analysis, elemental analysis, Fourier Transform InfraRed (FTIR) Spectroscopy, and Scanning Electron Microscopy (SEM). The characterization results indicated that the Fe(III) modified biochars possess favourable surface and chemical properties for phosphate

Table 6 Cost (US$) of $FeCl_3$ treated biochar production for treatment of stormwater runoff

Process	Time (hours)	Power rating of muffle (KW)	Power consumption (KWh)	Average electricity price per unit (₹ per KWh)	Cost of electricity (₹) (A)	$FeCl_3$ treated Biochar	Cost of $FeCl_3$ treatment (₹) (B)	Total cost of biochar per kg (₹) (A+B)	Cost of biochar per kg (US$)
Pyrolysis	1.0	2.0	2.0	6.0	12.0	Without treatment	0	12.0	0.14
						1%	23.2	35.2	0.42
						5%	116.0	128.0	1.54
						10%	232.0	244.0	2.93

removal. Phosphate removal from synthetic stormwater was found to be 29%, 49%, and 61% for an initial phosphate concentration of 5 mg/L using 1%, 5%, and 10% $FeCl_3$ treated biochar, respectively. However, the nitrate removal efficiency was found to be 9%, 12.4%, 14.6%, and 17.0% for B0%, B1%, B5%, and B10% respectively, at an initial nitrate concentration of 50 mg/L. This is possibly due to the preferential adsorption of phosphate instead of nitrate by Fe modified biochar, and the high initial concentration of nitrate. The removal efficiencies were found to reduce with increasing phosphate concentration, while no such trends were observed in the case of nitrate. The phosphate removal efficiency of modified biochar (B10%) for stormwater runoff collected from commercial, institutional, industrial, and road/highway areas of the Delhi region was 40.5%, 57.1%, 40.3%, and 59.1%, respectively, with the variations effected by the initial phosphate loading. A study of the kinetics of phosphate adsorption by B10% biochar showed that the adsorption of phosphate removal follows pseudo-second-order rather than pseudo-first-order kinetics. Both Langmuir and Freundlich isotherm models, which indicate mono-layer and multi-layered heterogeneous adsorption respectively, were found to be well fit for the phosphate adsorption by B10% biochar. The maximum adsorption capacity, obtained from Langmuir isotherm, was 11.57 mg/g for phosphate removal by B10%. Lastly, an analysis of production and modification costs of bamboo-biochar indicated that modification of biochar is economically feasible.

Author contributions SY: Conceptualization, execution, draft writing and editing. HP: Conceptualization, execution, draft writing and editing. AKH: Conceptualization, supervision, reviewing and editing. KRR: Conceptualization, supervision, reviewing and editing.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability Necessary data has been provided along with this manuscript.

Declarations

Ethical approval Not applicable.

Consent to Publish Authors provide the consent to publish the manuscript.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

Ayub A, Raza ZA, Majeed MI, Tariq MR, Irfan A (2020) Development of sustainable magnetic chitosan biosorbent beads for kinetic remediation of arsenic contaminated water. Int J Biol Macromol 163:603–617. https://doi.org/10.1016/j.ijbiomac.2020.06.287

Deng Y, Li M, Zhang Z, Liu Q, Jiang K et al (2021) Comparative study on characteristics and mechanism of phosphate adsorption on Mg/Al modified biochar. J Environ Chem Eng 9(2):105079. https://doi.org/10.1016/j.jece.2021.105079

Haritash AK, Sharma A, Bahel K (2015) The potential of Canna lily for wastewater treatment under Indian conditions. Int J Phytoremediation 17(10):999–1004. https://doi.org/10.1080/15226514.2014.1003790

Haritash AK, Dutta S, Sharma A (2017) Phosphate uptake and translocation in a tropical *Canna*-based constructed wetland. Ecol Process 6:12. https://doi.org/10.1186/s13717-017-0079-3

He C, Liu Z, Wu J, Pan X, Fang Z, Li J, Bryan BA (2021) Future global urban water scarcity and potential solutions. Nat Commun 12(1):1–11. https://doi.org/10.1038/s41467-021-25026-3

Hernandez-Mena LE, Pécoraa AA, Beraldob AL (2014) Slow pyrolysis of bamboo biomass: analysis of biochar properties. Chem Eng 37:115–120. https://doi.org/10.3303/CET1437020

Huang Y, Lee X, Grattieri M, Yuan M, Cai R, Macazo FC, Minteer SD (2020) Modified biochar for phosphate adsorption in environmentally relevant conditions. Chem Eng J 380:122375. https://doi.org/10.1016/j.cej.2019.122375

Jain N, Yadav S, Taneja S, Ray S, Haritash AK, Pipil H (2024) Phosphate removal from urban stormwater runoff using Canna lily and Cyperus alternifolius-based bioretention system. Sustainable Water Resour Manage 10(2):1–10. https://doi.org/10.1007/s40899-024-01076-5

Kleinman PJ, Srinivasan MS, Dell CJ, Schmidt JP, Sharpley AN, Bryant RB (2006) Role of rainfall intensity and hydrology in nutrient transport via surface runoff. J Environ Qual 35(4):1248–1259. https://doi.org/10.2134/jeq2006.0015

Kumar S, Masto RE, Ram LC, Sarkar P, George J, Selvi VA (2013) Biochar preparation from Parthenium hysterophorus and its potential use in soil application. Ecol Eng 55:67–72. https://doi.org/10.1016/j.ecoleng.2013.02.011

Li S, Jiang F, Lei T, Ren Z, Wang S, Yang X (2019) Phosphorus removal by in situ sprayed ferric chloride in Dianchi Lake: efficiency, stability, and mechanism. Process Saf Environ Prot 131:320–328. https://doi.org/10.1016/j.psep.2019.09.021

Ministry of Agriculture & Farmers Welfare (2022) Contribution of Agriculture Sector towards GDP. Agriculture has been the bright spot in the Economy despite COVID-19. https://www.pib.gov.in/PressReleasePage.aspx?PRID=1741942. Accessed on 14 February, 2022

Mukherjee A, Zimmerman AR, Harris W (2011) Surface chemistry variations among a series of laboratory-produced biochars. Geoderma 163:247–255. https://doi.org/10.1016/j.geoderma.2011.04.021

Nandakumar S, Pipil H, Ray S, Haritash AK (2019) Removal of phosphorous and nitrogen from wastewater in Brachiaria-based constructed wetland. Chemosphere 233:216–222. https://doi.org/10.1016/j.chemosphere.2019.05.240

Panwar NL, Pawar A (2020) Influence of activation conditions on the physicochemical properties of activated biochar: a review. Biomass Convers Biorefinery 1–23. https://doi.org/10.1007/s13399-020-00870-3

Pipil H, Haritash AK, Reddy KR (2021) Seasonal variability and kinetics of phosphate removal in a Phragmites-based engineered wetland. Rend Lincei Scienze Fis E Naturali 32(4):729–735. https://doi.org/10.1007/s12210-021-01017-w

Pipil H, Haritash AK, Reddy KR (2022a) Spatio-temporal variations of quality of rainwater and stormwater and treatment of stormwater runoff using sand–gravel filters: case study of Delhi, India. Rend Lincei Scienze Fis E Naturali 33:135–142. https://doi.org/10.1007/s12210-021-01038-5

Pipil H, Yadav S, Taneja S, Chawla H, Haritash AK, Reddy KR (2022b) Water Sensitive Urban Design (WSUD) for Treatment of Storm Water Runoff. In: Kanwar, V.S., Sharma, S.K., Prakasam, C. (eds) Proceedings of International Conference on Innovative Technologies for Clean and Sustainable Development (ICITCSD – 2021). Springer, Cham. https://doi.org/10.1007/978-3-030-93936-6_5

Pipil H, Yadav S, Taneja S, Chawla H, Haritash AK, Reddy KR (2023) Removal of phosphate from Stormwater Runoff using bench scale constructed Wetland. In: Yukselen-Aksoy Y, Reddy KR, Agnihotri AK (eds) Sustainable Earth and Beyond. EGRWSE 2022. Lecture Notes in Civil Engineering, vol 370. Springer, Singapore. https://doi.org/10.1007/978-981-99-4041-7_39

Qiu B, Duan F (2019) Synthesis of industrial solid wastes/biochar composites and their use for adsorption of phosphate: from surface properties to sorption mechanism. Colloids Surf a 571:86–93. https://doi.org/10.1016/j.colsurfa.2019.03.041

Reddy KR, Xie T, Dastgheibi S (2014) Evaluation of biochar as a potential filter media for the removal of mixed contaminants from urban storm water runoff. J Environ Eng 140(12):04014043. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000872

Romero-Gueiza MS, Tait S, Astals S, del Valle-Zermeno R, Martinez M, Mata- Alvarez J et al (2015) Reagent use efficiency with removal of nitrogen from pig slurry via struvite: a study on magnesium oxide and related by-products. Water Res 84:286–294. https://doi.org/10.1016/j.watres.2015.07.043

Sinha E, Michalak AM, Balaji V (2017) Eutrophication will increase during the 21st century as a result of precipitation changes. Science 357(6349):405–408. https://doi.org/10.1126/science.aan2409

Smith JS, Winston RJ, Tirpak RA, Wituszynski DM, Boening KM, Martin JF (2020) The seasonality of nutrients and sediment in residential stormwater runoff: implications for nutrient-sensitive waters. J Environ Manage 276:111248. https://doi.org/10.1016/j.jenvman.2020.111248

Sun Y, Gao B, Yao Y, Fang J, Zhang M et al (2014) Effects of feedstock type, production method, and pyrolysis temperature on biochar and hydrochar properties. Chem Eng J 240:574–578. https://doi.org/10.1016/j.cej.2013.10.081

Wang Z, Guo H, Shen F, Yang G, Zhang Y, Zeng Y et al (2015) Biochar produced from oak sawdust by Lanthanum (La)-involved pyrolysis for adsorption of ammonium (NH_4^+), nitrate (NO_3^-), and phosphate (PO_4^{3-}). Chemosphere 119:646–653. https://doi.org/10.1016/j.chemosphere.2014.07.084

Wang Z, Shen D, Shen F, Li T (2016) Phosphate adsorption on lanthanum loaded biochar. Chemosphere 150:1–7. https://doi.org/10.1016/j.chemosphere.2016.02.004

Weber K, Quicker P (2018) Properties of biochar. Fuel 217:240–261. https://doi.org/10.1016/j.fuel.2017.12.054

Yang YY, Lusk MG (2018) Nutrients in urban stormwater runoff: current state of the science and potential mitigation options. Curr Pollution Rep 4(2):112–127. https://doi.org/10.1007/s40726-018-0087-7

Zeng Z, Zhang SD, Li TQ et al (2013) Sorption of ammonium and phosphate from aqueous solution by biochar derived from phytoremediation plants. J Zhejiang Univ Sci B 14(12):1152–1161. https://doi.org/10.1631/jzus.B1300102

Zhang G, Liu H, Liu R, Qu J (2009) Removal of phosphate from water by a Fe–Mn binary oxide adsorbent. J Colloid Interface Sci 335(2):168–174. https://doi.org/10.1016/j.jcis.2009.03.019

Zhang L, Zhou Q, Liu J, Chang N, Wan L, Chen J (2012) Phosphate adsorption on lanthanum hydroxide-doped activated carbon fiber. Chem Eng J 185:160–167. https://doi.org/10.1016/j.cej.2012.01.066

Zhang M, Song G, Gelardi DL, Huang L, Khan E et al (2020) Evaluating biochar and its modifications for the removal of ammonium, nitrate, and phosphate in water. Water Res 186:116303. https://doi.org/10.1016/j.watres.2020.116303

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 13, Issue. 5, May 2024, pg.107 – 115

Fingerprint-Enabled ATM Network

Pragati Agrawal[1]; Rohan Saxena[2]; Saksham Agrawal[3]; R.K. Singh[4]

[1,2,3,4](Department of Mechanical Engineering & Delhi Technological University, India)

[1] pragati1234ag@gmail.com; [2] rohansaxena858@gmail.com;
[3] agrawalsaksham1323@gmail.com; [4] rohansaxena859@gmail.com

DOI: https://doi.org/10.47760/ijcsmc.2024.v13i05.011

Abstract: Large data collections are stored by people in the modern world utilizing computers. Data may be sent across the globe at a reasonable cost. ATM's have been in use for almost 50 years now since their introduction. Since its introduction, ATM's have made it possible to withdraw money without encountering any problems. In the banking industry, the ability of ATM's to be installed in any location is seen as exceptional since it allows customers the freedom to conduct transactions on their own. While the cash dispenser and visual screen serve as output devices, the card reader and keyboard serve as input devices. Many ATM's can work together on a common level through the host processor, which is owned by the bank. Authenticated Banks provide unique ATM PINS to their customers. The registered phone number is used to send these ATM PINS. if someone saw the PIN notice on the user's phone and knows the card data. They can then access the client's account in full and may be held accountable for any criminal activity that occurs. Criminals have developed a variety of schemes to steal money from ATM's. Due to strict security measures around the money compartment, they are unable to simply steal cash from an ATM; instead, they swipe the cards of several customers. It is far easier to obtain ATM card data than to breach the ATM's security. The practice of card skimming is one that con artists frequently employ. On the card per user space of ATM's, card skimmers are implemented. The data is stored in the memory of the card skimmer and is seen by an electrical circuit on a visually appealing portion of the card. After driving the skimmer out of the ATM, the criminals later duplicate this data. Card skimmers are designed to fit perfectly over each user's unique cards, making it easy for clients to become victims of them.

Keywords: OTP, Database, SQL, NOSQL, PIN, URL

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

I. INTRODUCTION

Our goal is to introduce a potent technique that uses fingerprint sensing to enhance the current state of ATM's. It is widely acknowledged that each person has a fingerprint that is distinct from everyone else. Thus, we will lessen the likelihood of thefts occurring in the present day, since the offenders can access systems through technological means and obtain all of the user's private information, but they will be prevented from completing the transaction if the fingerprint is not authenticated. In order to undertake operations only after a comprehensive validation of the user's fingerprint from our database, we are additionally incorporating One Time Passwords that are sent via the registered user's phone number and email address as ATM PIN. Criminals have developed a variety of schemes to steal money from ATMs. Due to strict security measures around the money compartment, they are unable to just take cash out of ATMs; instead, they swipe the cards of various customers. It is much easier to steal ATM card info than to breach directly into the ATM. The practice of card skimming is one that con artists frequently employ. On the card per user space of ATMs, card skimmers are implemented. The data is stored in the memory of the card skimmer and is seen by an electrical circuit on a visually appealing portion of the card. After driving the skimmer out of the ATM, the criminals later duplicate this data. Card skimmers are designed to fit over unique cards perfectly making it easy for customers to become victims of it. In the unlikely event that card spaces appear large or unrestricted, it may be a skimmer device. Additionally, fake key cushions that log all customer inputs are available online from obscure web commercial centers. These can be placed over the ATM's primary key cushion.

Then, thieves take advantage of these logged inputs to withdraw money from the client account. Additionally, thieves can place a tiny camera directly above the ATM keypad, which keeps recording customer inputs, including your PIN. Banks frequently advise their customers to shield the keyboard with their other hand when accessing the PIN. Occasionally, the entire front of the ATM is replaced with a false one that exactly mimics the original. It turns out to be challenging to determine whether an ATM has been compromised in the event of a phony ATM front. Money doesn't flow out of the fake front, and customer inputs are still being recorded.

II. OVERIEW OF FINGERPRINT AUTHENTICATION

- Step 1 Registration: The registration process involves first detecting the user's fingerprint, processing it into a digital image, and then purging any undesirable features from it. After post- processing, store the image in the database.
- Step 2 Authentication: To operate on a fingerprint, we used a fingerprint sensor. It extracted features from the fingerprint and compared them to N fingerprint templates stored in our database.
- Step 3 Confirmation: After a fingerprint was found to match a specific fingerprint in our database, we double-checked the match by comparing the fingerprint to the database once more.

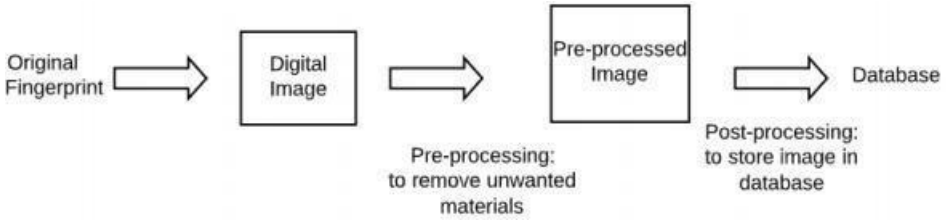


Fig. 1 Registration

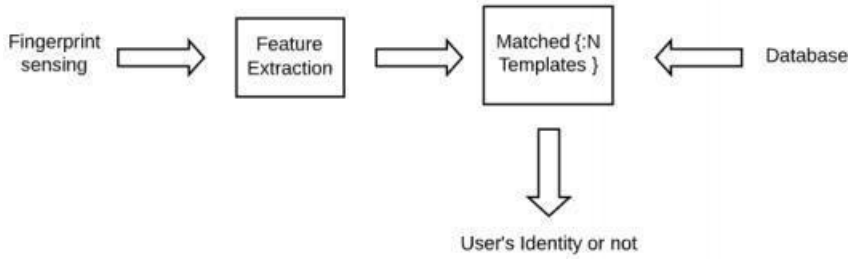


Fig. 2 Authentication

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

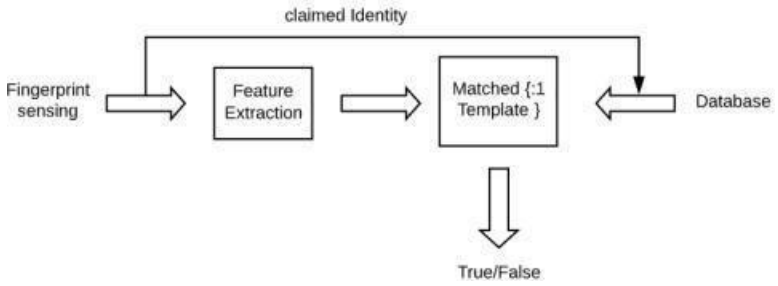


Fig.3 Confirmation

In addition to fingerprint recognition, the matching fingerprint customer's registered phone number, 99XXXXXXXX, and registered email address, customeruser@example.com, will receive three One Time Passwords. By all means, he can complete transactions if he can accurately enter the OTPs on or before the third OTP is transferred.

He will have to start the process over if the amount he inputs to be withdrawn is more than what is already in his bank account.

After the transaction is complete, he will get an email and a text message on his phone saying that the transaction on this account number, 56XXXXXX, was successful and that the remaining amount is what he needs to pay. Conversely, following the collapse ofthe third OTP, he will get a security breach alert email and be instructed to get in touch with the closest branch right away.

Face recognition problems are one way to counteract the existing state of affairs. As a result of biological ageing, the characteristics of a person's face become increasingly out of proportion from the last picture they had taken. For the device to remain productive, the user must periodically re- register their image, which could be extremely difficult for a bedridden patient or someone who has had their face burned. Comparing biometric IDs to more conventional and contemporary approaches has several advantages. Installing it is fairly easy, although it does require some time and effort to locate each finger and fingerprint scanner. Because of this, fingerprint recognition is regarded as one of the easiest biometric verification methods to use.

A webcam was mounted on the ATM, according to [23], and it records the clients' live photos as they pass by. After then, this taken image is compared to the other pictures in the database. An arbitrary password is generated by the contrast and sent to the client's registered phone when it is successful. The customer is then directed to input the identical password on the ATM's screen. The user can carry out transaction actions if they input correctly. However, there are two significant problems with this theory: 1) Unless prompt action is done, it will be impossible to stop theft from occurring if the mobile phone is lost or stolen. 2) The initial image was created because as we age, our face expressions tend to alter. There will be differences between the picture of our faces that was taken at that time and the most recent one.

According to the researchers, the idea of using the PIN as a secure password is not very suitable for the ATM's Palm Print Recognition and Confirmation System [4]. Thus, the idea of fingerprint authorization was established in order to increase security. According to extensive research, a simulator mimics the features of the ATM as it exists today. They compared the cervical recognition program's 89.43% rejection rate to 10.57% in their conclusion [4].

As a result, 53 out of 500 users who visit the updated ATMs after going through this verification procedure may experience issues. The percentage of rejections was increased to ten percent.

To eliminate errors in ATMs' alphanumeric certification, Lalzirtira implemented Graphical User Authentication. His research indicated that an image-integrated code would be easier to understand. He devised a password that would be simple for customers to commit to memory [5]. Although the pass-code that uses photographs to grant authorization in the ATM system has some benefits of its own, it was unsuccessful because of video recordings.

Dyno-password idea was another one that was proven to be true. Here, the bank sends the client a text message on the registered phone number, which contains the passcode. The ATM screen will be used to input this new passcode. The bank's database server will cross-check it once more. It follows that all that is required to withdraw cash from an ATM is the customer's PIN, card, and SIM. This kind of sensitive information can be readily obtained by anyone in the client's social circle. This idea thus provided opportunities for hackers [6].

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

III. SOFTWARE

A. *Mongo DB*

An open-source record database with excellent performance, high accessibility, and designed scaling is called MongoDB. To put it simply, Mongo DB is a database that is oriented towards records. It is an open-source product that was developed and supported by 10gen. MongoDB can be accessed for free under a general public licence, and it can also be accessed with a commercial licence from the creator. "A document-oriented database system that makes advantage of the NoSQL notion is called Mongo DB. It also offers automatic scaling, great performance, and high availability. This open- source application was created and is maintained by 10gen, a company that was founded in October 2007. MongoDB is a free database management tool offered under the General Public Licence (GPL) and Commercial licence issued by the producer. It was also planned for MongoDB to run on commodity systems. MongoDB is the database of choice for businesses of all sizes and in all sectors worldwide. A database in MongoDB can be thought of as a physical space containing various types of data. Every database has a live assortment of documents here on the record framework. A MongoDB server often houses a number of databases.

B. *Python Spyder*

Like Perl, Ruby, Scheme, or Java, Python is an object-oriented programming language that is powerful and easy to understand.

Among Python's noteworthy attributes are:

- Its elegant syntax makes your programme easier to read.
- The language is user-friendly, making it effortless to get your programme up and running. Because of this, Python is perfect for a variety of ad-hoc programming tasks, including developing prototypes, without sacrificing maintainability. It has an extensive standard library that supports a wide range of common programming tasks, including connecting to web servers, using regular expressions to search text, reading, and editing files, and testing brief code snippets with ease. Additionally, thereis the IDLE packaged development environment.
- Easily expandable with the addition of new modules written in compiled languages like C or C++.
- It is compatible with Mac OS X, Windows, Linux, and Unix, and there are unofficial builds for Android and iOS as well.
- Free software is defined in two ways. The cost of downloading, using, and incorporating Python into your programme is zero.
- Despite being protected by copyright, Python is freely modifiable and re-distributable due to its open- source licence.

C. *Tkinte*r

There are several solutions available in Python for creating a GUI (Graphical User Interface). Tkinter is the most widelyused approach of all the GUI methods. It is a typical Python interface to the Python-bundled Tk GUI toolkit. The quickest and most straightforward approach to construct GUI apps is with Python and Tkinter. Using Tkinter to create a GUI is a simple process.

Making a Tkinter application:

- Bringing in the Tkinter module
- The process involves creating the main window (container), adding any number of widgets, and then applying the event trigger to each widget.

D. *Visual Studio (Node JS)*

Node.js provides a framework for developing efficient and expandable server applications utilizing JavaScript. Acting as the runtime environment, Node.js is complemented by npm, which serves as the package manager for Node.js modules. Visual Studio Code comes pre-equipped with built-in support for both JavaScript and Typescript languages, along with debugging capabilities for Node.js. However, for executing Node.js applications, it's essential to have the Node.js runtime installed on your local machine. This runtime facilitates the execution of JavaScript code outside the browser environment, enabling the development of diverse server-side applications with ease and efficiency.

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

IV. PROCEDURE

The process via which fingerprints are recognized:

Biometric fingerprint technology is the most widely acknowledged security mechanism to be placed in the ATM. Compared to the current ATM security; it offers a considerably more secure initiative. Because it is so simple to install and use, the technology's ability to blend seamlessly with our modern lifestyle is astounding. It is praised as one of the most effective authentication methods as a result. Since the fingerprint data are not kept in a database, there is no way for the fingerprint to be misused.[7]

Two essential presumptions underpin the unique finger impression ID: peculiarity and in-variance. Constancy. According to in- variance, the characteristics of the finger impression don't alter throughout time. Peculiarity: suggests that the one-of-a-kind mark is remarkable and no two people's distinctive finger impressions are alike.

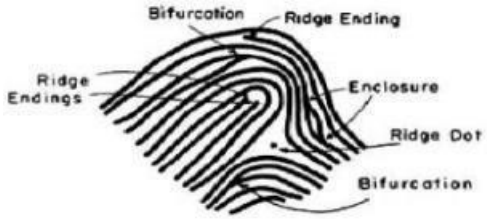


Fig. 4 Fingerprint characteristics [1]

Fig. 5 Fingerprint Types

- Distinctive biometric impression: The biometric's component example is its distinctive pattern. [21]
- Binarization: This refers to the process of converting a greyscale image into a binary image by determining the esteem value. [21]
- Block Filter: This technique effectively extracts minutiae by reducing the thickness of all ridgelines to a single pixel width. [21]
- Minutiae Extractruction: After minutiae are eliminated, details are deduced.[21]
- Minutiae Matching: This compares unique finger impression data. Additionally, the minutiae coordination layout is used. [21]
- Matching Score: This is the relationship used to determine how well the original fingerprint and template data collaborate. [21]

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

V. APPROACH FOR ATM MODIFICATION

They won't check for a one-time password from the bank using their registered phone number or email ID until their fingerprints have been carefully verified.

Three opportunities will be provided for them to write their accurate OTP in front of the ATM screen. In the event that there is enough money in their accounts, they will be able to withdraw money if they enter it correctly. However, if, despite our giving themthree One Time Passwords, they enter them wrong, we will send them an email saying, "Security Breach and Kindly Contact your nearest branch." API's will be used by us for sending and receiving obtain the customer's uniquely identifiable data.

The framework is introduced to carry out specified assignments, like email correlation, phone compatibility, validation of Python code, and so on. Each method is then reset to begin regulations. Because the intended technology includes fingerprint sensing, it differs from the current ATM system. It will start as soon as the customer walks into the ATM. First, it will be indicated if the person wishing to open an account with our bank is an existing customer or a new one. If the person is already a customer, they will be requested to move straight on to the procedure' fingerprint authentication step. Should they choose not to, they will be required to register with us by providing their phone number, email address, and fingerprint in our database.

The pictures of our encrypted database are as follows:

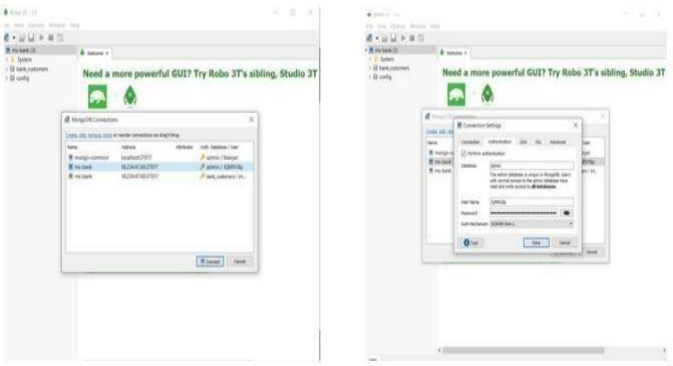


Fig. 6 Database with Password Protection

It indicates that the ability to access the database is restricted to the owner of the verified credentials, such as the admin password. Following that, fingerprints will be requested from both new and current customers in order to confirm their inclusion in our database. As long as their fingerprint scan yields a satisfactory confidence match with our database, they are free to proceed. Their ability to participate in the following phases will be restricted if the fingerprint authentication is unsuccessful. In order to prevent fraudulent activities, it is necessary to ensure that the amount of sensitive information that hackers can readily trace, such as card numbers, personal information security, and mobile numbers, will decrease to the point where people feel safe and secure.

On the successful verification of the fingerprint, they will be asked to enter the first OTP send to their registered Email Address and Phone Number. If they enter it correctly, they will be asked to enter the amount they want to withdraw. If the amount entered is less than the balance present in the bank account, they will successfully withdraw it. On the backend side, we will update the remaining balance in our database. If they fail to enter their first OTP accurately, then a second OTP will be sent to their Email Address and Phone Number again with following the same consequences. On the third OTP, they will have only one chance to enter it fairly. If they failed to do so, then from our side, a security breach mail and a text message will be sent to their registered Email Address and Phone Number, and they will be locked out to proceed for the transaction of money. They have to start over with the whole process again. We are using Python GUI Tkinter as the interface to take the input from the user. We are using API's and Mongo DB to extract and deliver the data of customers on the requirement. We have deployed an online URL for registering using WIFI with us: www.msbank.co.in. This URL is responsible for generating APIs.

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

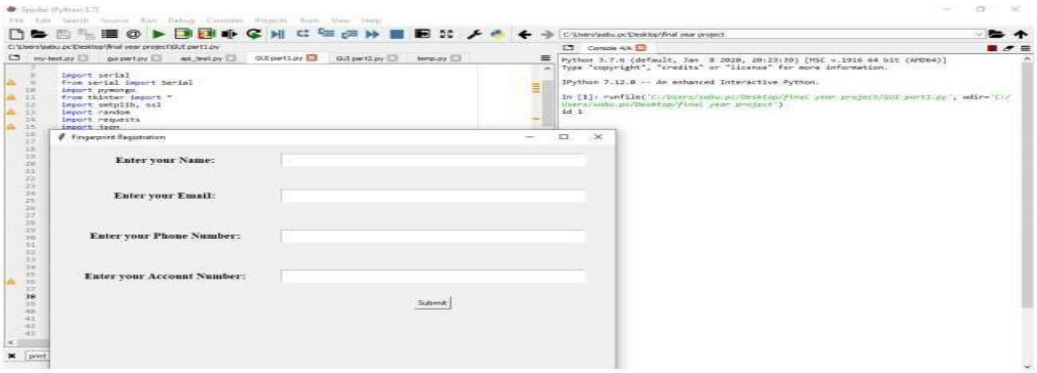


Fig. 7 Illustrates a graphical user interface (GUI) designed in Python for data entry purposes.

VI. RESULTS AND DISCUSSIONS

In this context, we submit the technical input of our research project.

A. *Database*

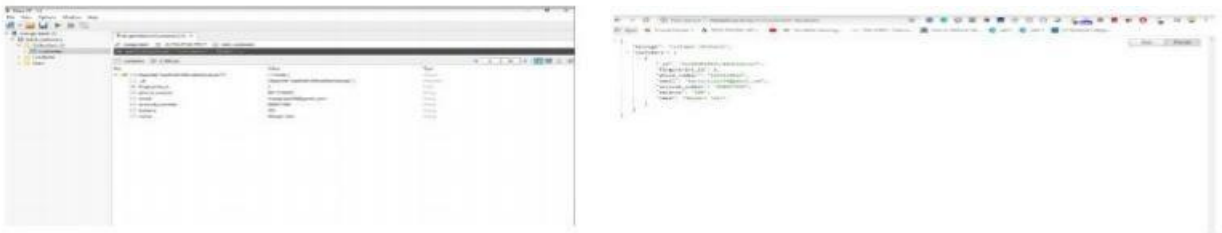


Fig. 8 Our initial database, Mongo DB, is shown in image a) APIs are shown in image b) It displays our database of active MS Bank customers.

B. *Updating the existing database by adding a new entry*

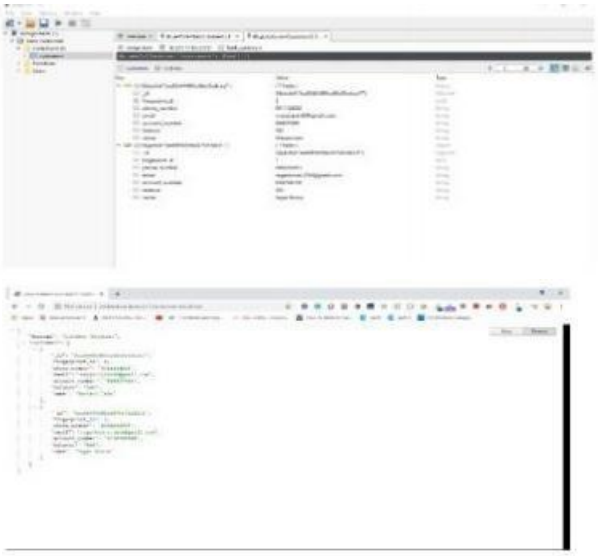


Fig. 9 Shows how the API's are updated with the new user's database in Image b and Mongo DB in Image a. It demonstrates that the new data has been successfully added to our database and API's.

C. *Successfully entering the OTP*

According to the amount in his or her bank account, it indicates that every registered customer who enters the OTP matching from the texts and emails is permitted to perform transactions.

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

Fig. 10 Effective Fingerprint Verification (A new GUI displays requesting the first OTP to be provided here after the newly registered user's fingerprint is successfully verified.)

D. Incorrect entry of OTP

This is the situation that arises when a registered customer enters the OTP incorrectly for the third time. They will consequently receive an email titled "Security alert, kindly contact your nearest branch as soon as possible."

Fig. 11 First OTP is incorrect The client inputs the initial OTP incorrectly.(The GUI presents a warning saying "Wrong otp 1" when the customer enters the first OTP, which doesn't match the one sent to the registered Email and Phone No. After that, in order to get the second OTP, the consumer must select "Click to Receive OTP.")

VII. CONCLUSIONS

The existing research on enhancing ATM security through fingerprint authentication, as outlined in the Literature Synopsis, revealed certain inconsistencies. Consequently, there was a pressing need to enhance ATM security with a more effective and cost- efficient solution. The proposed approach involves integrating fingerprint authentication, but with a deviation in the method of delivering OTP's compared to previous studies. OTP's are now sent via two channels: the client's registered email address and phone number. The methodology entails offering new clients the option to register with the bank through a Python GUI (ATM Screen). For registered clients, the ATM Fingerprint Sensor first verifies their biometrics. Subsequently, an OTP is sent to their registered contact details. In the event of an incorrect entry for the first OTP, the client has two additional attempts to input it accurately. If the third attempt fails, the client (or suspected hacker) is barred from conducting any transactions, and a Security Theft Alert email is dispatched. Additionally, the URL www.msbank.co.in facilitates online registrations for clients and the creation of API's for securely storing data of new and enrolled clients. This system ensures that clients can still access OTP's even in scenarios where Wi-Fi is unavailable, by utilizing mobile OTP's, or retrieve lost OTP's from their logged email address. In conclusion, the research findings validate the feasibility of this proposed approach.

ACKNOWLEDGEMENT

With immense pleasure and a deep sense of gratitude, we wish to express our sincere thanks to our supervisor Prof. R.K. Singh, DTU, without his motivation and continuous encouragement, this research would not have been successfully completed. We are grateful to the Chancellor of Delhi Technological University for motivating us to carry out research in the Delhi Technological University and also for providing us with infrastructural facilities and many other resources needed for my research. We like to acknowledge the support rendered by our friends and faculties in several ways throughout our research work. We wish to extend my profound sense of gratitude to our parents for all the sacrifices they made during our research and also providing us with moral support and encouragement whenever required.

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

REFERENCES

[1] F. A. Afsar, M. Arif and M. Hussain, National Conference on Emerging Technologies, 2004, "Fingerprint Identification and Verification System using Minutiae Matching,", pp.141-146.

[2] Hmadri nath moulick, Joyjit Patra, Arun kanti Manna, International Journal of Engineering Inventions, Volume 3, Issue 1, August 2013, "Computer Assisted and Contour Detection in Medical Imaging Using Fuzzy Logic", pp-1-7.

[3] Kavita Hooda, International Journal of Scientific and Research Publications, Volume 6, Issue 4, April 2016. "ATM Security", pp-1.

[4] Sanjay, S. G., International Journal Of Engineering And Computer Science, 2014, "ATM Transaction Security System Using Biometric Palm Print Recognition and Transaction Confirmation System", pp-5332-5335.

[5] Lalzirtira, National Institute of Technology Rourkela Mtech Thesis, 2013, "Graphical User Authentication, India".

[6] Anand, D. A., Dinesh, G. & Naveen, H. D., International Journal of Communication and Computer Technologies (IJCCT), 2013. "A Reliable ATM Protocol and Comparative Analysis on Various Parameters with other ATM Protocols", pp. 192- 197.

[7] Steve Furber, ARM System-on-Chip Architecture, Second Edition, 2000, ISBN 0-201- 67519-6.

[8] Ratha, N., Connell, J. & Bolle, R., IBM Systems Journal, Volume 40, No. 3, 2001, "Enhancing Security and Privacy inBiometrics-based Authentication Systems", pp. 614-634.

[9] Petrlic, Ronald, and Christoph Sorge, IET Information Security, Volume 8, Issue 2, 2013, "Establishing user trust in automated teller machine integrity", pp-132-139.

[10] Myo, N., International Conference on Education Technology and Computer, 2009, "Fingerprint Identification Based on the Model of the Outer Layers of Polygon Subtraction", pp-201 – 204.

[11] Journal for Research, Volume 2, Issue 12, February 2017

[12] Ms. Archana S. Shinde and Prof. Varsha Bendre, 2015 International Conference on Computing Communication Control and Automation, 2015, "An Embedded Fingerprint Authentication System", pp-45.

[13] Yuliang He, Jie Tian, Xiping Luo, Tanghui Zhang, Pattern Recognition, Letters 24, 2003, "Image enhancement and minutiaematching in fingerprint verification", pp-1349-1360.

[14] Wei Wang, Jianwei Li, Feifei Huang, Hailiang Feng, Pattern Recognition, Letters 29, 2008, "Design and implementation of Log-Gabor filter in fingerprint image enhancement", pp-301- 308.

[15] Lin Hong, Wan Yifei, Anil Jain, IEEE Transactions on Pattern Analysis and Machine intelligence, Volume 20, Issue 8, 1998,"Fingerprint image enhancement: algorithm and performance evaluation", pp- 777-789.

[16] Der Chin Chen, Biometric Systems, Design and Applications, 2011, "Portable Biometric System of High Sensitivity Absorption Detection".

[17] Subhra Mazumdar, Venkata Dhulipala, San Diego, "Biometric Security Using Finger Print Recognition".

[18] Anil K. Jain, Jianjiang Feng, Karthik Nandakumar, IEEE Computer Society 2010, 0018- 9162/10, "Fingerprint Matching", pp. 36-44.

[19] Sowmya Ravikumar, Sandhya Vaidyanathan, B. Thamotharan, S. Ramakrishnan, International Journal of Engineering and Technology, Volume 5, No. 3, 2013, "A new business model for ATM transaction security using fingerprint recognition".

[20] Jun Zhou, Guangda Sua, Chun hongJiang, Neurocomputing, 70, 2007, "A face and fingerprint identity authentication system based on multiroute detection" pp-922-931.

[21] Mithun Dutta, Kangkhita Keam Psyche and Shamima Yasmin, American Journal of Engineering Research (AJER), Volume 6, Issue 8, August 2017, "ATM Transaction Security Using Fingerprint Recognition", pp-41-45.

[22] Zhao, F., Tang, X., The Journal of the Pattern Recognition Society, Volume 40, Issue 4, April2007, "Preprocessing and postprocessing for skeleton-based fingerprint minutiae extraction", pp-1270-1281.

[23] Frimpong Twum, Kofi Nti, Michael Asante, International Journal of Science and Engineering Applications, Volume 5, Issue 3, May 2016, "Improving Security Levels In Automatic Teller Machines (ATM) Using Multifactor Authentication", pp-503.

© 2024, IJCSMC All Rights Reserved, ZAIN Publications, Fridhemsgatan 62, 112 46, Stockholm, Sweden

Check for updates

Received: 25 November 2023 | Revised: 11 May 2024 | Accepted: 20 May 2024

DOI: 10.1002/cpe.8194

RESEARCH ARTICLE

WILEY

FLOLSTM: Fuzzy logic-driven optimized LSTM for improved malicious traffic detection in hypervisor environments

Anumukonda Naga Seshu Kumar[1] | **Rajesh Kumar Yadav**[1] | **Nallanthighal Srinivasa Raghava**[2]

[1]Department of Computer Science and Engineering, Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, India

[2]Department of Electronics and Communication Engineering, Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, India

Correspondence
Anumukonda Naga Seshu Kumar, Department of Computer Science and Engineering, Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, India.
Email: seshu1203@gmail.com

Summary

In the ever-evolving realm of cloud computing, the challenge of intrusion detection has grown increasingly intricate and vital. With the proliferation of cyber-attacks and the widespread use of virtualized environments, there is a pressing need for network security solutions that are not only innovative and robust but also easily comprehensible. These solutions must possess the ability to effectively detect malicious activities, provide visibility into network operations, adapt to changing requirements, and promptly alert stakeholders to any suspicious behavior. In this study, we introduce a groundbreaking approach known as fuzzy logic-driven optimized long short-term memory (LSTM)—FLOLSTM, specifically designed for hypervisor-based environments. By integrating fuzzy logic with an optimized LSTM neural network, FLOLSTM aims to significantly enhance the detection of attack traffic within hypervisor networks by combining the interpretability and uncertainty management capabilities of fuzzy logic with the temporal pattern recognition prowess of LSTM. The research methodology involves meticulous data collection from hypervisor monitors, followed by rigorous cleaning and preprocessing to enhance data reliability. Subsequently, the preprocessed data is input into the FLOLSTM classifier to identify malicious activities. Furthermore, the performance of the LSTM is fine-tuned using the waterwheel plant optimization (WPO) algorithm. Experimental evaluations compare the efficiency of the proposed FLOLSTM with existing techniques across various metrics including accuracy, recall, precision, F-measure, specificity, false-positive rate (FPR), and false-negative rate (FNR). Overall, the FLOLSTM model represents a significant advancement in intrusion detection for cloud environments, offering a potent blend of interpretability, accuracy, and efficiency. Its superior performance underscores its capacity to enhance network security and effectively mitigate cyber threats in dynamic and virtualized settings, thereby making a substantial impact on the field of cyber security.

KEYWORDS

attack detection, fuzzy logic, hypervisor, LSTM, malicious, optimization, traffic detection, waterwheel plant

 © 2024 John Wiley & Sons Ltd.

1 | INTRODUCTION

Cloud computing systems allow users to access a pool of computing resources on demand on a recurring basis. Cloud service providers manage the resources and users may access them through the Internet.[1] As a result of its multi-tenant nature, cloud computing systems can accommodate multiple virtual machines (VMs) on a single physical infrastructure. Since the world and the internet are deeply integrated, the network framework is always vulnerable to attack. Due to the fact that cloud computing systems are exposed to multiple security threats, cloud computing architecture is characterized by virtualization technology. As a result of this technology, several virtual machines are able to share physical machine resources, such as the CPU, memory, I/O, and network interfaces, between them. For cloud service providers (CSPs), protecting guest operating systems from advanced sophisticated attacks is one of the greatest challenges they face.[2–4]

A key component of virtualization[5] is the hypervisor, which enables resource sharing among virtual machines. Furthermore, it exposes CSP to a higher security risk as a result. As soon as someone gains access to the hypervisor, they are completely in control of the virtual environment as soon as they obtain it. In the context of cloud computing, an intrusion may come from a number of places, such as virtual machines,[6] virtual networks,[7] or malicious hypervisors.[8] Due to their ease of rental and direct CSP connection to the outside world, virtual machines are a favorite target for attackers.[9] Attacker uses a variety of techniques to find virtual machine vulnerabilities so that they can be used as a backdoor to the virtual machine. These days, the area of interest is the technical problem of identifying these attacks. Intruders violate all of the core privacy requirements of secrecy, integrity, accessibility, and denial of service when they violate the trust imposed by the Internet.[11]

Intrusion Detection and Prevention System is a crucial component for safeguarding the entire virtual system from various threats and attacks.[10] This system should be able to identify malware, analyze logs, check files for integrity, analyze incoming traffic, and respond actively, among other things. It should also have the ability to identify both known and unknown assaults. In the absence of prevention, detection alone is insufficient to protect the virtual environment. Numerous methods have been put out in recent years to identify anomalies due to the rise in security concerns and the massive volume of communication.[12] These methods have a disadvantage in that the system must constantly update its knowledge of the attack types when new ones appear. Moreover, these systems cannot detect precise changes within the cloud hosting environment due to their narrow field of view. For the purpose of detecting malicious traffic in hypervisor systems, we provide a Fuzzy Logic-driven Optimized LSTM in order to get around these problems. Fuzzy logic's strengths in handling uncertainty and an optimized LSTM's ability to capture temporal trends are combined in this classifier to effectively detect malevolent conduct.

The main contributions of the paper are as follow:

- We present the integration of fuzzy logic with an optimized LSTM neural network to create a hybrid model that leverages the strengths of both approaches, enhancing the detection of attack traffic within hypervisor networks.
- Initially, data is collected from hypervisor monitors, then cleaned and preprocessed to improve data reliability.
- After that, the preprocessed data is sent into the FLOLSTM classifier to identify the detection of malicious activities.
- To optimize LSTM model, we introduce the waterwheel plant optimization (WPO) algorithm for enhancing both efficiency and effectiveness.
- We perform a comparative evaluation of the proposed FLOLSTM against existing techniques, assessing performance metrics such as precision, accuracy, F-measure, recall, specificity, FNR, and FPR.

The complete paper is structured as follows: The study on privacy intrusion detection in hypervisor networks, which had done by various researchers, is compiled in Section 2. Section 3 provides specifics on the proposed methodology for our research. Section 4 illustrates and explains the consequences and results of several strategies. The conclusions and recommendations for further research are presented in Sections 5.

2 | LITERATURE REVIEW

This section discusses several studies related to intrusion detection in hypervisor networks. These studies focus on various methods for detecting and preventing malicious attacks, as well as on developing algorithms that can be used to analyze network security data.

Pilli et al.[13] proposed Program Semantic-Aware Intrusion Detection at Network and Hypervisor Layer (PSI-Net Visor), a method based on out-VM monitoring, to detect assaults at the network or virtualization levels in cloud computing. PSI-Net Visor performed network monitoring by employing a behavior-based intrusion detection technique at the connection layer of a centralized Cloud Network Server. PSI-Net Visor employed Virtual Machine Introspection libraries using software break point injection to acquire process execution traces from the hypervisor. Depth first search (DFS) was also utilized to generate program semantics using execution trace flow of control graphs. It used dynamic evaluation and machine learning methods to comprehend the nature of anomalies, making it resistant to obfuscation and encryption-based assaults. This method takes a long time to train larger databases of different VM instances.

Garg and colleagues[14] developed a network intrusion detection system model that accurately predicts anomalous attacks with low misclassification error. They started by deleting duplicate entries and one that contained an incident using a local host IP. They employed neural networks,

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

a naive Bayes classifier, a decision tree, an SVM, an extreme gradient boosted tree, an AdaBoost classifier, a random forest classifier, and logistic regression models in this technique. When running cross-validation on each model, random states were employed to avoid over-fitting. The results showed that the extreme gradient boosted tree separated out from the crowd and performed admirably in contrast to the other models in this study. However, the dataset employed in this work lacks real-time data that simulates recent threats.

Singh and Ranga[15] presented a method for detecting assaults and intrusions in cloud computing systems using an ensemble learning methodology. The goal of this work is to implement an efficient network-based intrusion detection system using an ensemble-based machine learning method with four types of classifiers such as boosted tree, subspace discriminant, bagged tree, along with RUSBooted combined using a voting scheme. An aggregated prediction can be made at the end as a result of the voting algorithm's integration into the framework. CICIDS 2017 and CloudSim were used to simulate and test the proposed model. It has been observed that the Intrusion Detection System does not perform well when network packets have been encrypted.

Abbas and colleagues[16] proposed a new hybridized technique for intrusion detection and attack classification. The dataset was first preprocessed with the data transformation methodology and the min–max approach. The random forest recursive feature reduction method was then used in a second stage to select optimal features which improve the model's performance. Then, to detect intrusions, fine Gaussian support vector machine (FGSVM) kinds were utilized, and adaptive neuro-fuzzy system (ANFIS) types were employed to categorize probe, R2U, U2R, and DDOS attacks. As a result, FGSVM and ANFIS could be strong solutions with large efficiency gains. However, there are some limitations to this approach. However, the time required identifying optimal features may be too long, leading to slow intrusion detection.

Niazi and Faheem[17] developed a Bayesian game-theoretic detection of invasions method in smart grids for hypervisor-based software defined networks. The tactical interaction among a hypervisor observing its vSDN controllers with the origin of new traffic requests made by switches infiltrated by a DDoS attacker was modeled as a noncooperative dynamic Bayesian intrusion detection scenario in this work. This game model allowed a hypervisor to optimally deploy the resources it had to track guest vSDN controllers. The simulation findings demonstrated that while this game model allows a hypervisor to raise the chance of identifying distributed attacks and limit false positives, it also has a high monitoring cost.

Suri and colleagues[18] used WLI-FCM and Artificial Neural Networks to create a Cloud Intrusion Detection System for the Internet. A hybrid technique that merged the WLI-FCM clustering method with the back propagation artificial neural network was utilized to enhance the accuracy of detection of the cloud intrusion detection method. This model is made up of three steps. In the first step, a fuzzy clustering module is utilized to partition big datasets into small clusters in order to increase ANN capacity for learning. A variety of Artificial Intelligence modules are trained based on cluster values in the second step. Finally, the final result was derived by combining the results of the different Artificial Intelligence modules from the second step. This method achieved better performance but the complexity of the model can be difficult to manage and may require additional resources to maintain.

Jaber and Rehman[19] created a system for intrusion detection that uses a fuzzy c-means clustering (FCM) method in conjunction with support vector machines (SVMs). Hybrid clustering has been used for the creation of the hypervisor inspector, combining SVM and FCM. In the first phase, the FCM clustering module was used to separate big datasets into small clusters so that SVMs could learn effectively. This process enhanced the performance of the SVM through the fuzzy clustering module. The second phase involved training different SVM modules based on the cluster values. In the third module, the hypervisor inspector's results are combined. As a result, while this technique was capable of detecting diverse threats having high detection accuracy along with low false alarm rates, it was not designed for real-time applications.

Rawashdeh et al.[20] demonstrated an anomalous intrusion detection system in the hypervisor layer as a method for inhibiting DDoS between virtual machines. They created IDS using soft computing approaches, as well as PSO using a feed forward neural network. An evolving neural network was utilized to classify the communication transferred among virtual machines. Particle swarm optimization was combined using neural networks using the evolutionary neural network. By using particle swarm optimization, the neural network weights were selected in such a way that they could achieve high levels of accuracy in both classification and detection. However, it is restricted to only two types of DDoS attacks in this dataset.

The existing models developed by researchers have shown promising results in enhancing security measures and detection capabilities in network environments. The IMFL-IDSCS technique[21] introduces fuzzy logic-based IDS for cloud security, achieving high detection efficiency with accuracy. A hybrid intrusion detection model[22] combines signature-based and anomaly-based detection methods, demonstrating high detection rates across various datasets. Cloud-based IDS with ML techniques[23] explores performance evaluation and enhancement opportunities, proposing a system with improved performance. A hybrid deep learning approach[24] achieves high accuracy rates in network security, particularly effective in preventing DoS/DDoS attacks in SDN environments. Additionally, the CoAP-IoT Anomaly Detection Dataset[25] offers a comprehensive dataset for anomaly detection in CoAP traffic, utilizing shallow machine learning techniques for validation and analysis.

3 | PROPOSED METHODOLOGY

This research introduces a novel methodology that combines fuzzy logic and LSTM recurrent neural networks to detect attack traffic in hypervisor network. Figure 1 shows the overall structure for proposed methodology. The process begins with data collection from hypervisor monitors, followed by data cleaning and preprocessing. Then the preprocessed data is sent into the FLOLSTM classifier, an innovative fusion of fuzzy logic

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

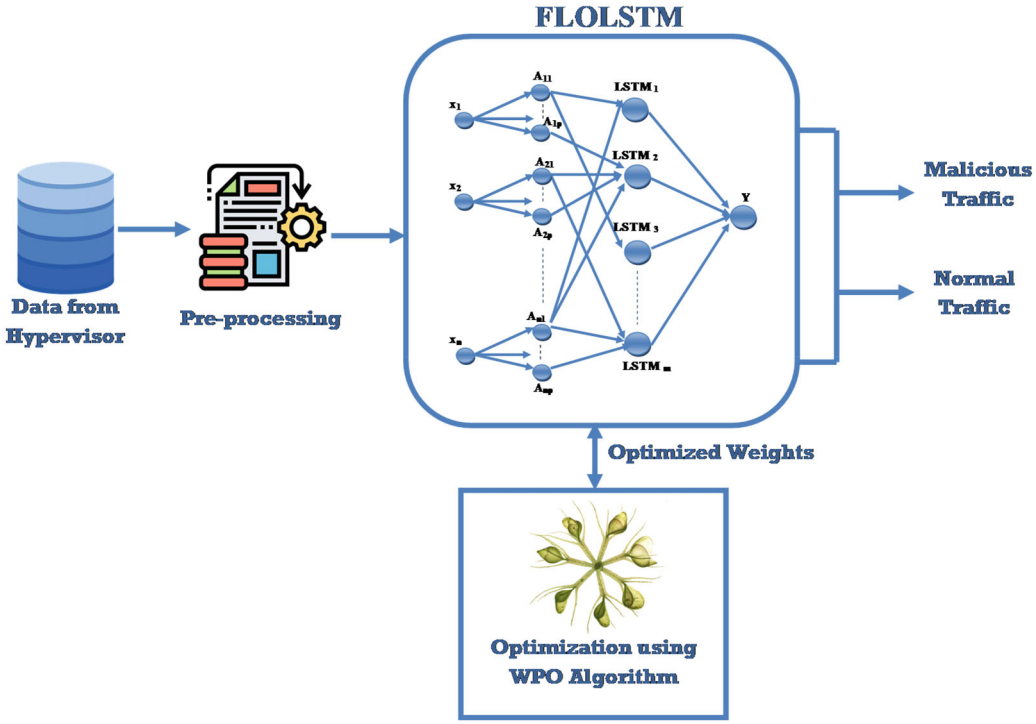


FIGURE 1 General structure of proposed methodology.

and optimized LSTM, designed to identify the detection of malicious activities. Moreover, the performance of the LSTM is optimized by using WPO algorithm. Fuzzy logic lends interpretability and adeptly manages uncertainties, while optimized LSTM excels at capturing temporal patterns. This classifier combines the strengths of fuzzy logic for handling uncertainty and optimized LSTM for capturing temporal patterns, enabling effective detection of malicious behavior.

3.1 | Data collection

In this work, we gather data from the hypervisor monitors, which are responsible for monitoring the traffic flow in the hypervisor network. These monitors continuously capture network-related metrics and events, such as "size" (packet size), "fragmented" (whether the packet is fragmented or not), "seqNumber" (sequence number of the packet), and "ackNumber" (acknowledgment number of the packet).The data collected here would typically consist of a mix of normal traffic and instances of attack traffic. Attack traffic includes any malicious or abnormal network activities that can potentially harm the system or compromise its security.

3.2 | Data cleaning and preprocessing

In this step, we clean and prepare the collected data from the hypervisor monitors for further analysis and modeling. Data collected from hypervisors may not always be in a perfect state. It may contain errors, contradictions, or missing values. Data cleaning entails operations such as deleting duplicates, dealing with missing data, and fixing errors in order to guarantee that the information remains accurate and reliable. After cleaning, the data might still need additional processing to be useful for further analysis. Preprocessing can involve tasks such as normalization and scaling data to a standard range and converting data into a suitable format.

- **Null Value Removal:** We check the dataset for any missing or null values. If any are found, we apply appropriate techniques (e.g., imputation) or simply remove the affected data points.
- **Remove Duplicate Values:** Sometimes, data collected from monitors might contain duplicate entries due to network latency or other reasons. Removing duplicates ensures that our model does not learn from redundant information.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

- **Label Encoder:** Categorical data (if present) is frequently converted into numerical format to aid in the training of machine learning models using approaches such as Label Encoding.
- **MinMax Scalar:** To make sure that all features are of a same scale, feature scaling is used. The MinMaxScaler scales numerical features to a range typically between 0 and 1.
- **Random Over Sampler:** If we encounter class imbalance in the dataset (i.e., when the number of samples in one class outweighs the other significantly), we use techniques like Random Oversampling to create synthetic examples of the minority class. It assists in balancing the class distribution and keeps the model from favoring the majority class.

3.3 | FLOLSTM model implementation

After preprocessing, the preprocessed data is sent into FLOLSTM for further classification. FLOLSTM is a novel methodology that combines fuzzy logic and optimized LSTM recurrent neural networks to detect malicious traffic in hypervisor network. Fuzzy modeling using LSTM cells is a technique that combines the principles of fuzzy logic with the capabilities of optimized LSTM neural networks. Fuzzy logic is a mathematical framework for dealing with uncertainty and imprecision, while LSTM is a type of recurrent neural network that is well-suited for modeling sequences and time series data.

Fuzzy logic plays a key role in interpreting linguistic variables and defining rules for decision-making based on expert knowledge or data-driven approaches. By fuzzifying input variables, applying fuzzy inference rules, and defuzzifying outputs, fuzzy logic helps the model make sense of complex data and generate actionable insights. On the other hand, LSTM networks are adept at capturing temporal dependencies and patterns in sequential data, making them well-suited for detecting anomalies and suspicious activities in intrusion detection.

By employing LSTM's ability to learn long-term dependencies, the FLOLSTM model can effectively analyze network traffic over time and identify potential security threats. The synergy between fuzzy logic and LSTM in the FLOLSTM model enhances decision-making by combining the interpretability of fuzzy rules with the predictive power of LSTM. Fuzzy logic provides a transparent framework for reasoning and rule-based decision-making, while LSTM offers the capability to learn from historical data and make predictions based on sequential patterns. Combining these two approaches can be useful in situations where we need to make predictions or decisions based on uncertain or imprecise information.

The FLOLSTM system, depicted in Figure 2, consists of four key layers, each with a specific role.

- **Input Organization:** The first layer of FLOLSTM arranges incoming data, gathered from hypervisor monitors, ensuring it is ready for processing.
- **Fuzzification:** Fuzzification is the process of converting crisp input signals into fuzzy sets using membership functions. In the second layer of FLOLSTM, data is transformed into fuzzy representations, allowing the system to handle uncertainty more effectively. The Gaussian membership functions define the degree of membership of each input signal to each fuzzy set.
- **IF and THEN Calculation:** The third layer of FLOLSTM is the core of the system, where it evaluates fuzzy rules, determining the degree of rule satisfaction and deriving corresponding actions. The fuzzy logic operation "AND" is used to calculate the value of the IF and THEN portions.
- **Output Estimation:** In the final layer, the FLOLSTM system combines fuzzy logic's interpretability with optimized LSTM's ability to capture temporal patterns, resulting in precise predictions and decisions, making it valuable for tasks like malicious activity detection in a hypervisor network.

In discrete time, a nonlinear function can be used to depict a system in this way:

$$y(k) = \varphi[X(k)] \tag{1}$$

Further, in addition, we define the state vector as $X(k)$: where $\varphi[\cdot]$ is a nonlinear difference equation of unknown variables.

$$X(k) = \left[y(k-1), \ldots y(k-a_y), x(k), \ldots, x(k-a_x)\right]^T = [x_1. \ldots x_n]^T \tag{2}$$

Assuming that $x(k)$ and $y(k)$ represent the system's input and output values, respectively, a_y represents the delayed output value, a_x represents the delayed input value, and n represents the total amount of items x_n in $X(k)$.

To represent the system, we employ fuzzy IF-THEN rules identical to those used in a normal TS fuzzy system, with the pth rule having:

$$\text{Rule}_p : \text{If } x_1(k) \text{ is } A_{1p} \,\&\, x_2(k) \text{ is } A_{2p} \,\&\, \ldots\ldots \,\&\, x_n(k) \text{ is } A_{np}, \text{ then } h_p(k) = \rho_p(k).$$

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

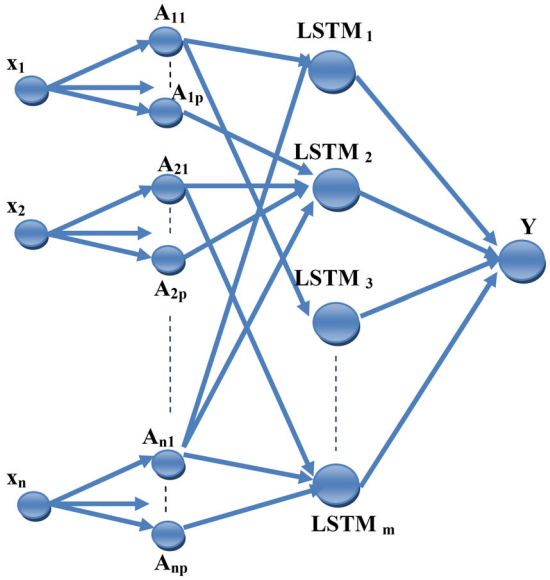


FIGURE 2 FLOLSTM classifier.

Assuming that each fuzzy rule's subsequent component is represented by the function $\rho_p(k)$, where $h_p(k)$ is the estimation function. The fuzzy sets for each x_n in (2)'s fuzzification is the sets A_{np}, for $n = 1 \ldots k$.

The following describes the membership functions associated with each A_{np}.

$$\mu_{A_{np},x_n}(k) = \exp\left(-\frac{\left(x_n(k) - \varsigma_{np}\right)^2}{2v_{np}}\right) \tag{3}$$

The value of the width is $v_{np} \in \Re+$, while its center is $\varsigma_{np} \in \Re$ in this Gaussian function. The T-norm yields the influence of every input element to the assumption part (IF part) of a fuzzy rule in (3) for the final assessment of the system.

$$z_p(k) = \prod_{n=1}^{k} \mu_{A_{np},x_n}(k) \tag{4}$$

It is possible to describe the values of each member of (4) in each fuzzy set more broadly using vectorial methods:

$$\varsigma_n = \exp\left[\left(X(k) - \chi_n\right)^2 \otimes \left(-\frac{1}{2}\gamma_n\right)\right] \tag{5}$$

Using $\chi_n, \gamma_n \in \Re^n$ as the corresponding width and center vectors for $\varsigma_n \in \Re^n$. In the fuzzy set A_n, every element of $x_n(k)$ is represented by the vector ς_n, and $\otimes$, the operator for the element-to-element product in vectors is denoted by gf. The fuzzy-network's parameters can be adjusted with the help of this representation. $h_p(k)$ is a representation of the following part (THEN part) of one fuzzy rule in (3).

The fuzzy system (3) is defuzzified using p rules to yield the estimate of (1).

$$\hat{y}(k) = \frac{\sum_{p=1}^{n} z_p h_p(k)}{\sum_{p=1}^{n} z_p} = \sum_{p=1}^{n} \bar{z}_p h_p(k) \tag{6}$$

Where,

$$\bar{z}_p = z_p/(z_1 + z_2 + \ldots \ldots + z_n) \tag{7}$$

Additionally, Figure 3 depicts the LSTM cell in the subsequent section. The "gate" mechanism is used by the cells to process data and allows useful information to flow through their structure. Compared to a traditional RNN, this cell can handle both short- and long-term data dependencies more effectively. The cells can be arranged as an array or as a network to cooperate with one another.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

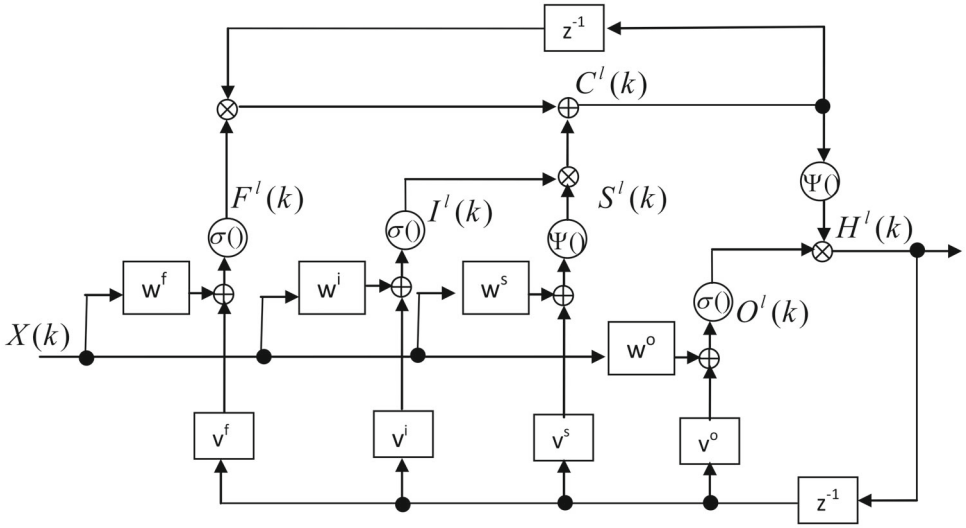


FIGURE 3 LSTM cell for the THEN part.

There are several stages in the LSTM network, which are described as follows:

$$F^l(k) = \sigma\left(w^f X(k) + u^f H(k-1)\right) \tag{8}$$

$$I^l(k) = \sigma\left(w^i X(k) + u^i H(k-1)\right) \tag{9}$$

$$S^l(k) = \psi\left(w^s X(k) + u^s H(k-1)\right) \tag{10}$$

$$C^l(k) = F^l(k) \otimes C^l(k-1) + I^l(k) \otimes S^l(k) \tag{11}$$

$$O^l(k) = \sigma(w^o X(k) + u^o H(k-1) \tag{12}$$

$$H^l(k) = O^l(k) \otimes \psi\left(C^l(k)\right) \tag{13}$$

In this case, the parts of the network represented by the, $I^l(k), F^l(k) S^l(k), O^l(k), C^l(k), H^l(k) \in \Re^p$ are the internal input fitness, internal state fitness, internal input, output fitness, internal state, and output of the LSTM network, respectively. As diagonal matrices, $u^f, u^i, u^s, u^o \in \Re^p$ as vectors, $w^f, w^i, w^s, w^o \in \Re^{p \times n}$; $u^f, u^i, u^s, u^o \in \Re^{p \times p}$ depending on the situation, are the synaptic weights. Since $X(k) \in \Re^n$ is the input in Equation (2), the functions $\sigma(\cdot)$ and $\psi(\cdot)$ are the hyperbolic tangent function and sigmoid function, respectively.

The fuzzy system's output is derived from (6).

$$\hat{y}(k) = Z_F H^l(k) \tag{14}$$

The elements of the IF sections are represented by $Z_F \in \Re^n$, while the THEN parts are represented by $H^l(k) = \left[h_1(k), \ldots, h_p(k)\right]^T$. Both the total number of fuzzy rules and the amount of LSTM cells can be expressed as $p = k^n$ for a single estimation; for multiple estimations, the definition is $p = m(k^n)$, where m represents the total number of estimations ($\hat{y}(k) = \Re^m$), as was previously mentioned (6). Based on fuzzy systems function approximation theories, the detected nonlinear process (1) can be expressed as follows:

$$Y(k) = Z_F(w^*) H^l(w^*) + \mu(k) \tag{15}$$

The unmodeled dynamic $\mu(k)$ can be minimized by the unknown weights, represented by w^*. In this stage, this system effectively combines the strengths of fuzzy logic, adept at managing uncertainty and interpreting data, with the temporal pattern capturing capabilities of LSTM cells. As a result, the system is well-equipped to make informed and precise predictions, decisions, or assessments, making it especially valuable in complex scenarios such as the detection of malicious activities in a hypervisor network.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3.3.1 | Optimization with the WPO algorithm

To enhance the performance of the FLOLSTM, its weights are optimized by using WPO algorithm. This process fine-tunes the weights, aligning them for optimal efficiency and effectiveness. It empowers the FLOLSTM to capture intricate temporal patterns and expertly manage uncertainties in data sequences. The main source of inspiration for the basic idea of WPO is the method used by waterwheel plants to discover, catch, and subsequently move its prey of insects to a more convenient position prior eating it.

The tiny, translucent flytrap-like traps which waterwheel plants carry are attached to broad petioles. In order to guard against damage or accidental activation from other water plants, a ring of hair-like bristles encircles the trap. The outside edges of the trap are coated in many hook-like teeth this, like flytrap teeth, interlock when the trap shuts over its prey. In Figure 4, the waterwheel plant is shown.

The steps involved in WPO are explained as follows:

Initialization

The population, also known as the Q matrix, is initially initialized by Waterwheel plant. Consequently, every waterwheel stands for a vector-based solution. Here, weights of the FLOLSTM (w^f, w^i, w^s, w^o) in the network should be taken into account as the solutions in this study. The initial solution format is

$$Q_i = \{q_1, q_2, \ldots, q_n\} \tag{16}$$

Here, Q_i is the population of the solution, n represents the variables, then

$$q_1 = \left\{w_1^f, w_1^i, w_1^s, w_1^o\right\} \tag{17}$$

Fitness function

In each iteration, the fitness value of a solution is meticulously calculated, and the overall fitness function is determined by selecting the maximum accuracy score achieved among these solutions and it can be represented as;

$$fitness = Max(Accuracy) \tag{18}$$

Updation phase

Phase 1: Position identification and hunting of insects. Waterwheels have an excellent sense of smell, which allows them to track and detect pests with ease. When an insect gets close to the waterwheel, its strikes and follows its prey. Through the integration of the waterwheel's assault on the insect, the WPO improves its capacity to investigate the ideal area and prevents itself from becoming stuck in local optima.

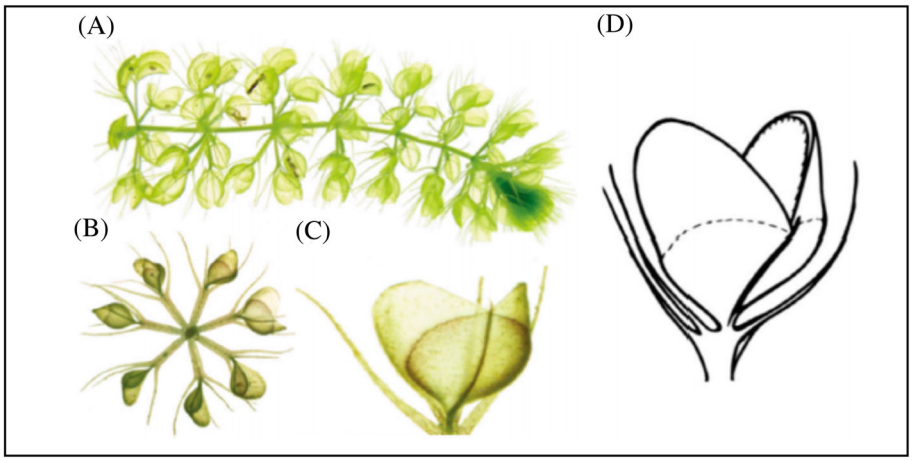


FIGURE 4 Image of a waterwheel plant (A) Lateral view of a shot floating freely and equipped with many traps. (B) Front view with both closed and open traps. (C) A solitary open trap. (D) Schematic illustration of an exposed trap.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

An equation is used to calculate the waterwheel's new location as it gets closer to the bug. The prior site is abandoned if moving the waterwheel to this alternative position increases the value of the objective function.

$$G = r_1 \cdot (Q(t) + 2E) \tag{19}$$

$$Q(t+1) = Q(t) + G \cdot (2E + r_2) \tag{20}$$

where the random variables r_1 and r_2 have values in the range of [0, 2] to [0, 1]. Additionally, E is an exponential variable having values in the range of [0, 1], and G is a vector that represents the circumference of the circle where the waterwheel plant would search for possible places. If the solution does not change shortly after three iterations, Equation (21) can be used to shift the placement of the waterwheel.

$$Q(t+1) = Gaussian(\mu_Q, \sigma) + r_1\left(\frac{Q(t) + 2E}{G}\right) \tag{21}$$

Phase 2: Position identification and hunting of insects. Waterwheels capture and transport insects to feeding tubes. WWPA uses this simulated behavior during local search to enhance its exploitation capability, enabling it to find solutions that are near to those previously discovered. By simulating the insect transport mechanism, a little shift in the waterwheel's location inside the search space is introduced. The WWPA designers assign each waterwheel in the population a random location in order to replicate the natural activity of waterwheels. If the objective function value of the waterwheel is higher at this new site, it will be transferred there and replace its old placement, as indicated by the equations that follow:

$$G = r_3 \cdot \left(EQ_{best}(t) + r_3 Q(t)\right) \tag{22}$$

$$Q(t+1) = Q(t) + EG \tag{23}$$

where the best solution is P_{best}, the current solution at iteration t is $P(t)$, and r_3 is a random variable having values in the range [0, 2].

Like in the exploration phase, the next mutation is applied to prevent being stuck in specific local minima if the solution fails to enhance after three repetitions.

$$Q(t+1) = (r_1 + E)\sin\left(\frac{B}{D}\theta\right) \tag{24}$$

where the random variables F and C have values between [−5, 5]. Equation 10 also shows that the value of K falls exponentially.

$$E = \left(1 + \frac{2 * t^2}{T_{\max}} + F\right) \tag{25}$$

Termination

The population update process persists until the algorithm's last iteration. At the conclusion of the WPO run, the output is the optimal solution, which encapsulates the optimized FLOLSTM model's weight configuration, reflecting the maximum attainable accuracy score.

Pseudo-code of WPO

Input: Weight parameter of FLOLSTM
Output: Optimized weight parameter of FLOLSTM

Initialize the weight parameter of FLOLSTM, the maximum number of iterations (T_{max}) and objective function f_n, the positions for n waterwheel plants, among other parameters.
Calculate the fitness f_n;
Find the optimal position Q_{best};
INITIALIZE t = 1;
While $t \leq T_{max}$ **Do**
  For ($i = 1: i < n + 1$) **Do**
  If ($r < 0.5$) **Then**
    Evaluate the waterwheel plant search space with equations (19) and (20);

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

If the solution remains unchanged for three iterations, then adjust the location using equation (21);
End If
Else
Utilize the existing solutions to determine the optimal solution using equations (22) and (23);
If the solution remains unchanged for three iterations, then adjust the location using equation (24);
End If
End If
End For
Reduce the value of E exponentially by equation (25);
Update the parameters;
Calculate the f_n for each position;
Identify the optimal position Q_{best};
$t = t + 1$;
End While
Return the optimal solution Q_{best}.

Therefore, FLOLSTM weight parameters are optimized through the WPO algorithm. FLOLSTM performs exceptionally well in classifying input datasets using these optimized weights, effectively distinguishing normal network traffic from malicious traffic. In this manner, the model can make informed and accurate decisions regarding the nature of the incoming information.

For an illustration

Our input dataset consists of various network-related metrics and events, such as, "size (x_1)" "fragmented (x_2)," "seqNumber (x_3)," and "ackNumber (x_4)." This each input having fuzzy set which has three sets of linguistic variables "small," "medium," and "large" which can be represented as A_{np}.

Layer 1: Each of this input attributes x_1, x_2, x_3, and x_4 is mapped to a fuzzy set, which in this case comprises three linguistic variables. For example, x_1 having fuzzy sets of $\{A_{11}, A_{12}, A_{13}\}$ and x_2 having fuzzy sets of$\{A_{21}, A_{22}, A_{23}\}$, and so on.

$$\text{i.e.}, x_1 : \{A_{11}, A_{12}, A_{13}\} \text{ where}, A_{11} \rightarrow \text{small}, A_{12} \rightarrow \text{medium}, A_{13} \rightarrow \text{large}.$$

Layer 2: In this layer, the crisp numerical values of the input attributes, such as x_1, x_2, x_3, and x_4 are transformed into fuzzy sets based on the linguistic variables. Fuzzification involves quantifying the degree to which each data point belongs to these fuzzy sets, effectively capturing the uncertainty and imprecision in the data.

Layer 3: In this layer, the system evaluates the "IF" part of fuzzy rules for each input attribute. In parallel, the "THEN" part of fuzzy rules is determined, representing the action or decision that follows the fuzzy rules. For example:

$$\text{RULE} : \text{IF } x_1 \text{ is } A_{11} \& x_2 \text{ is } A_{23} \& x_3 \text{ is } A_{32} \& x_4 \text{ is } A_{41} \text{ THEN } h_p(k) = \rho_p(k).$$

This part synthesizes the linguistic terms and membership functions to produce numerical outcomes.

Layer 4: The final layer involves the estimation of the overall system behavior based on the results of the previous layers whether they represent normal or malicious network traffic. If the output is −1 then it signifies normal data; else if it is 1 then designates the data as malicious.

4 | RESULTS AND DISCUSSIONS

To evaluate the effectiveness of our proposed framework, we discuss the results of the proposed framework and compare it with existing frameworks. The Python language was utilized to implement the system. The experiments were conducted within an environment comprising an Intel Core i5 processor running at 2.21 GHz, 8 GB of RAM, operating on the Windows 10 platform. The ISOT Cloud IDS (ISOT CID) dataset is used here which is a cloud intrusion detection dataset that contains over 8 TB of data collected in a real cloud environment. The dataset includes network traffic at VM and hypervisor levels, regular traffic, activities, and multiple attack scenarios. Table 1 represents the parameter settings of our experiment.

Sample output window

Figure 5 displays the graphical user interface (GUI) when the data is categorized as benign, and Figure 6 displays the same GUI when the data is classed as malicious. The metric values shown next to each figure represent the accuracy, recall, F-measure, precision, FPR, and FNR of the GUI when classifying the data as either benign or malicious.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Parameter settings.

Parameter	Values
Train:Test	80:20
Learning rate	0.001
Epoch	100
Input activation	Relu
Output activation	Softmax
Loss	Sparse categorical cross entropy
Input size	(1, 10)
Output	2
Batch size	2
Optimizer	Adam
Maximum number of iterations	100
r_1	[0,1]
r_2	[0,1]
r_3	[0,1]
Number of search agents	10
Number of runs	20

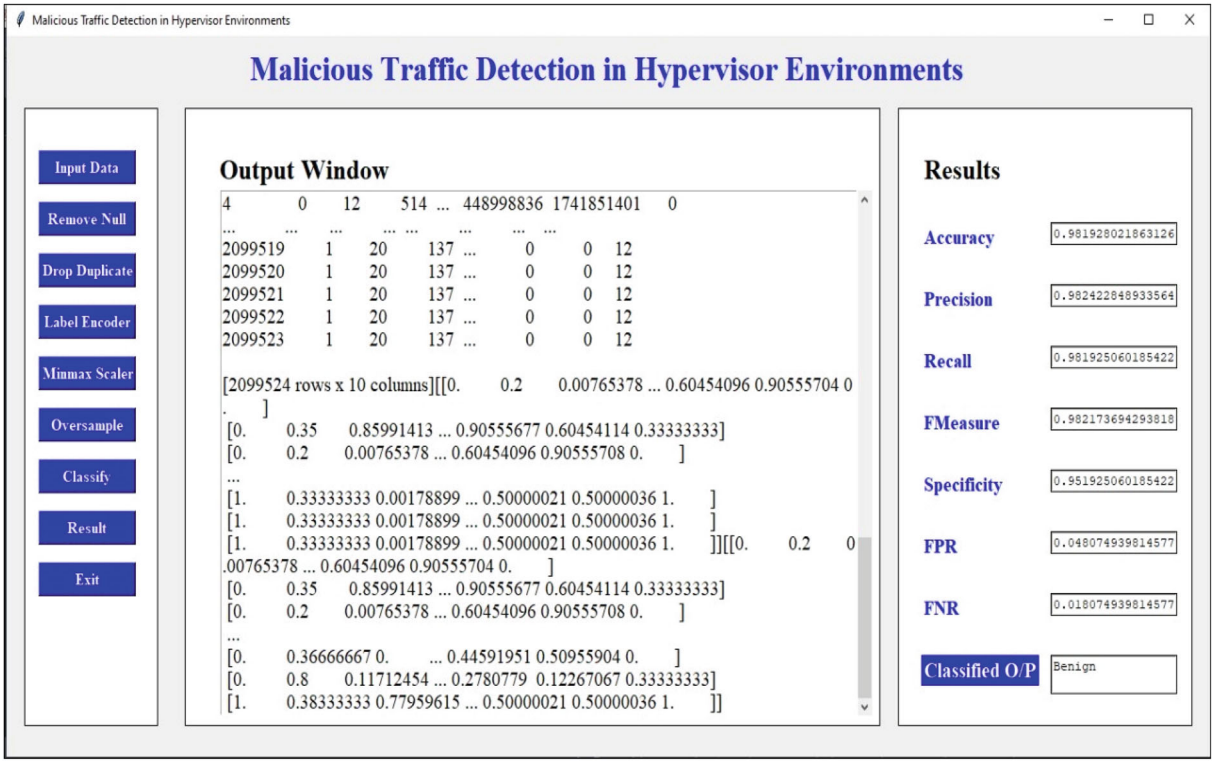


FIGURE 5 Sample output window of benign data classification.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

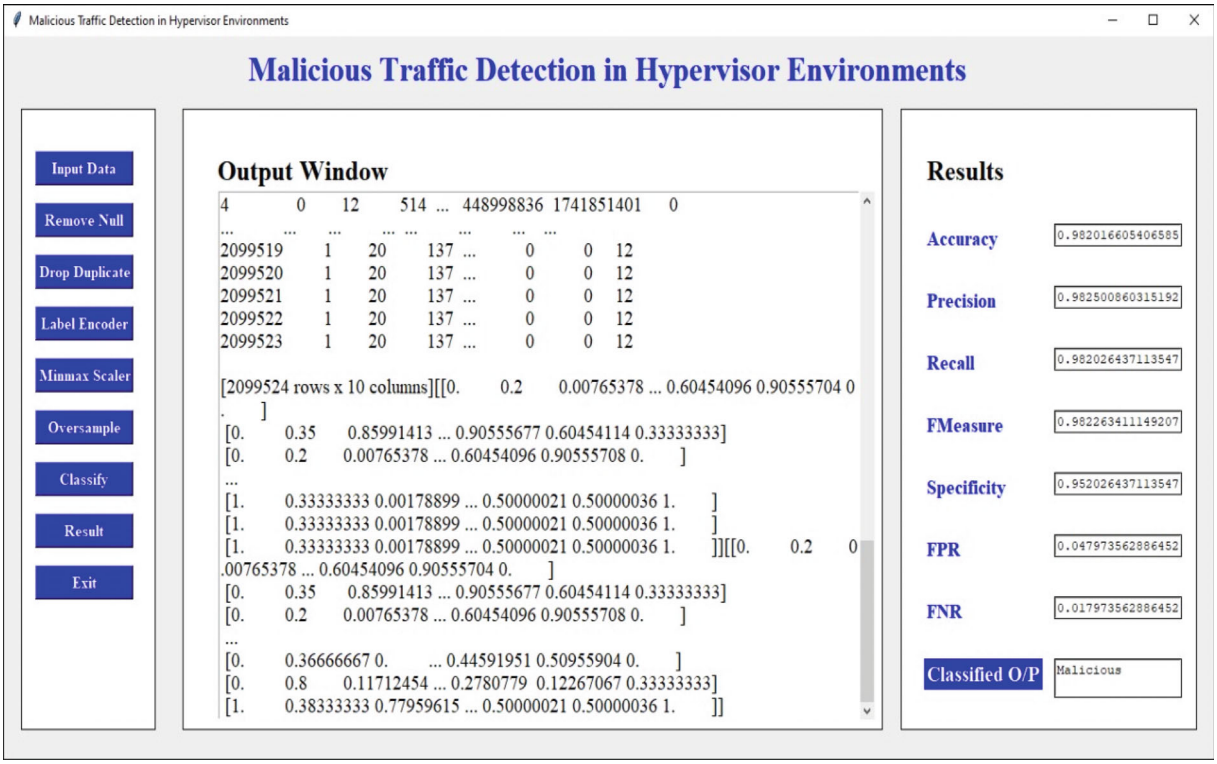


FIGURE 6 Sample output window of malicious data classification.

4.1 | Temporal analysis: Proposed FLOLSTM versus existing methods

In this research, we analyzed the hypervisor data in 4 phases that is Phase 1 (Day 1), Phase 2 (Day 2), Phase 3 (Day 3), and Phase 4 (Day 4) in order to observe how these metrics change over time or remain consistent over a period of time. We found that the proposed FLOLSTM model was able to predict the hypervisor attacks more accurately than the other approaches. We also observed a consistent trend in the data over the 4 days, indicating that the proposed FLOLSTM model is more reliable than the other approaches.

4.1.1 | Phase 1 analysis

Figure 7 showcases a robust performance of our proposed FLOLSTM model in differentiating between benign and malicious instances on Day 1. With 419,334 instances correctly identified as benign (True Positives), and an impressively low count of only five benign instances misclassified as malicious (False Positives), the FLOLSTM model displays high precision in its predictions. Notably, there are no instances of benign data incorrectly labeled as malicious (False Negatives), indicating an accurate identification of benign data. Moreover, the FLOLSTM model successfully identifies 418,809 malicious instances (True Negatives), demonstrating its proficiency in pinpointing harmful data.

The ROC curve of our proposed method in Day 1 can be seen in Figure 8. The ROC curve measures the accuracy of the model, and the higher the curve, the more accurate the model is. Our proposed FLOLSTM method has a ROC curve that is higher than other models, which indicates its accuracy.

4.1.2 | Phase 2 analysis

As shown in Figure 9, our FLOLSTM model performed well on Day 2 in identifying benign and malicious instances. Only 57,740 benign instances were misclassified as malicious (False Positives) in the FLOLSTM model, indicating high precision in its predictions. In addition, 15,633 benign data items were mistakenly classified as malicious (False Negatives), indicating accurate characterization of benign data. In addition, FLOLSTM is capable of identifying 620,239 malicious instances (True Negatives), proving its capability to identify malicious data.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

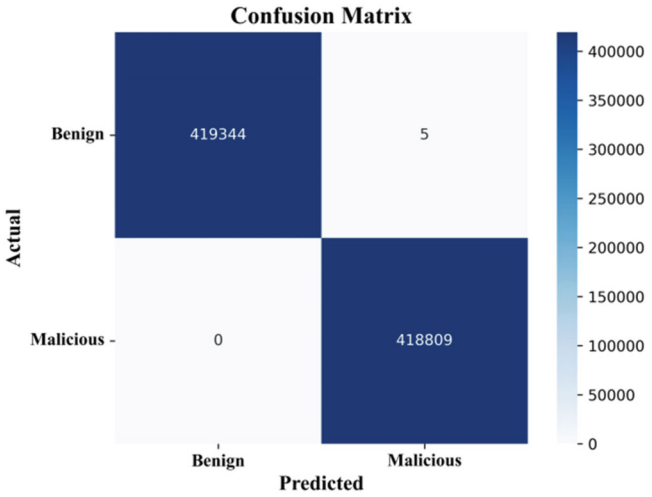


FIGURE 7 Confusion matrix of proposed FLOLSTM on Day 1.

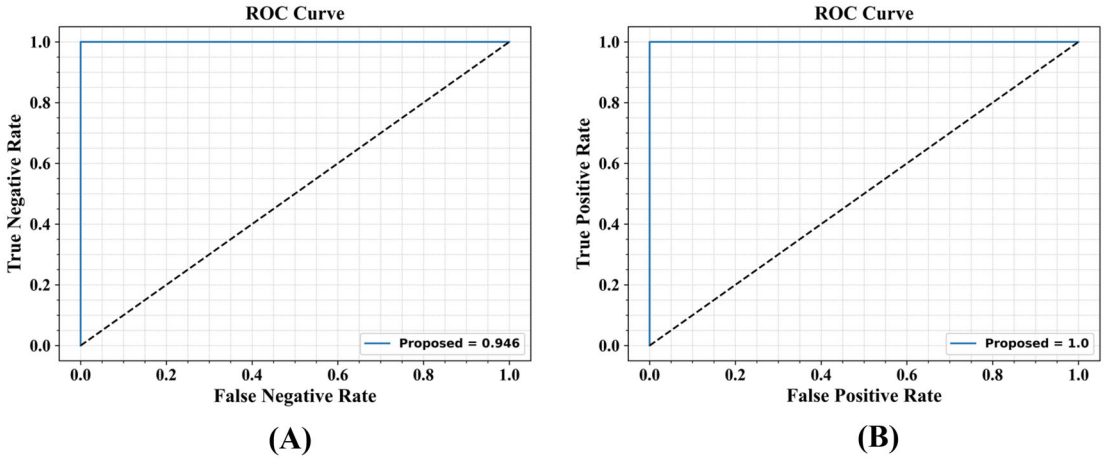


FIGURE 8 ROC curve of proposed FLOLSTM on Day 1 (A) TNR versus FNR (B) TPR versus FPR.

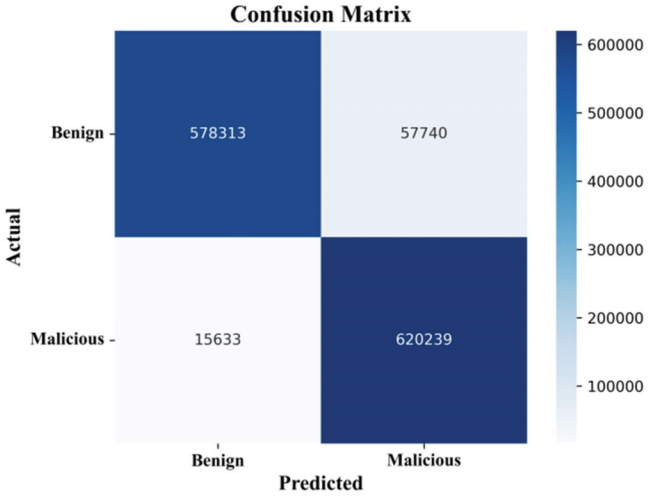


FIGURE 9 Confusion matrix of proposed FLOLSTM on Day 2.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Our proposed method's ROC curve in Day 2 is shown in Figure 10. The ROC curve measures how well our model can distinguish between true and false predictions. Our proposed method achieved a high AUC value, which indicates that it is highly accurate in predicting the outcome of each day.

4.1.3 | Phase 3 analysis

Based on Day 3 results, Figure 11 shows a powerful performance for our proposed FLOLSTM model in distinguishing benign from malicious instances. FLOLSTM accurately identified 270,653 benign instances (True Positives), and misclassified only seven benign instances as malicious (False Positives). There are 27 instances of benign data mistakenly labeled as malicious (False Negatives), indicating accurate identification of benign data. Furthermore, FLOLSTM identifies 270,981 malicious instances (True Negatives), demonstrating its ability to identify harmful data.

The ROC curve demonstrates the performance of the proposed model in Day 3, which has been shown in Figure 12. It shows that the proposed model was able to correctly classify the samples with an overall high accuracy.

4.1.4 | Phase 4 analysis

Our proposed FLOLSTM model differentiates benign from malicious instances on Day 4 as shown in Figure 13. It accurately identified 312,386 benign instances (True Positives), and identified only 3975 benign instances misclassified as malicious (False Positives). There are 5082 false

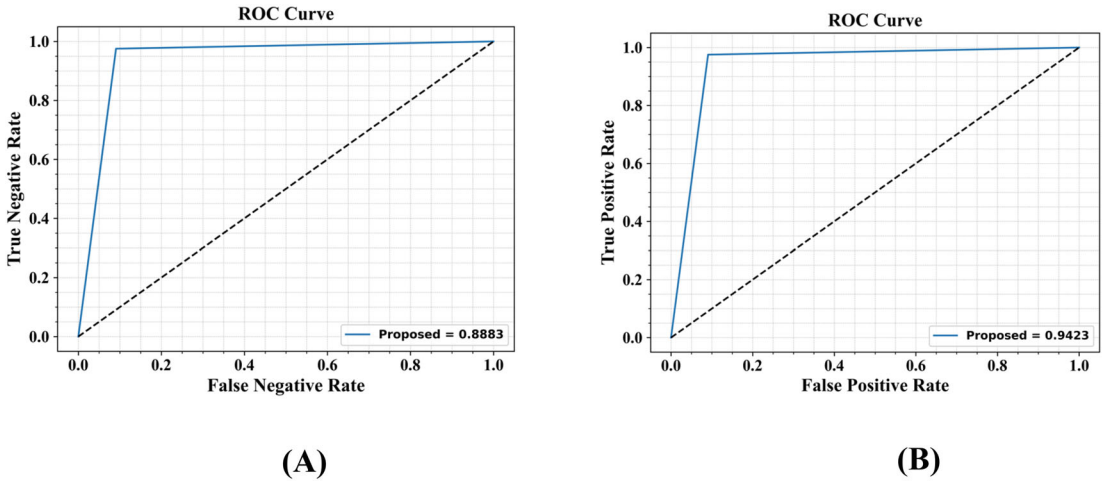


FIGURE 10 ROC curve of proposed FLOLSTM on Day 2 (A) TNR versus FNR (B) TPR versus FPR.

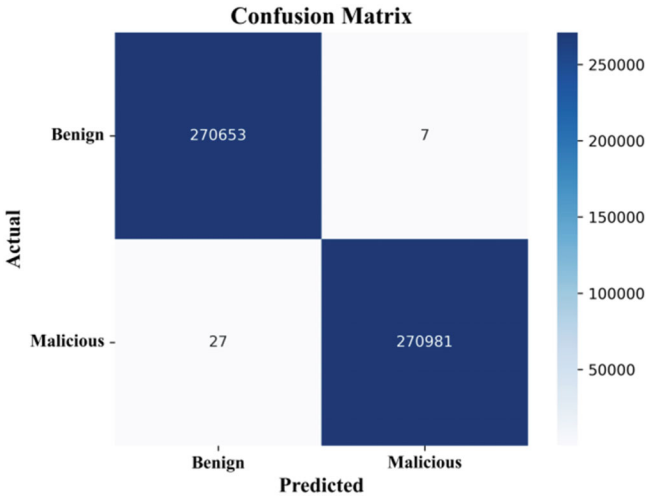


FIGURE 11 Confusion matrix of proposed FLOLSTM on Day 3.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

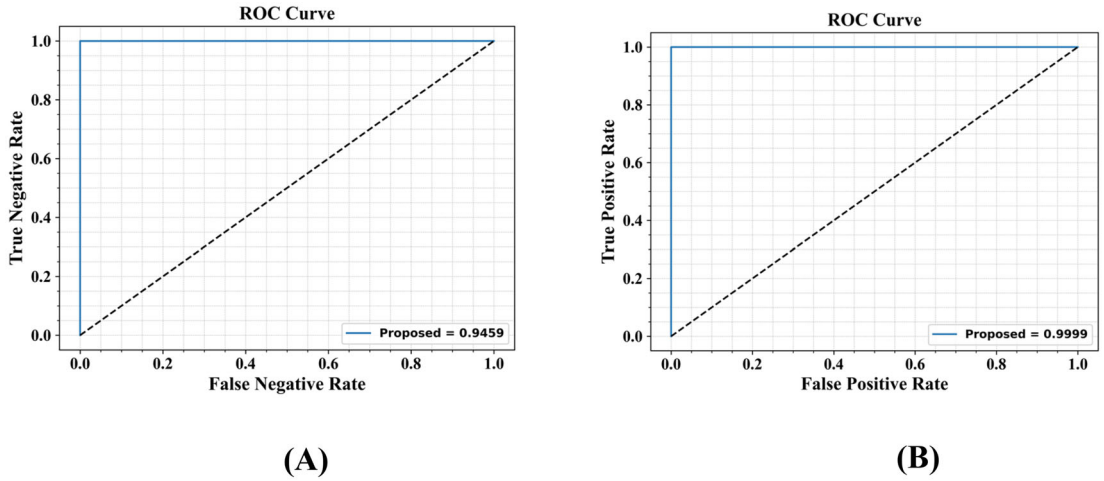


FIGURE 12 ROC curve of proposed FLOLSTM on Day 3 (A) TNR versus FNR (B) TPR versus FPR.

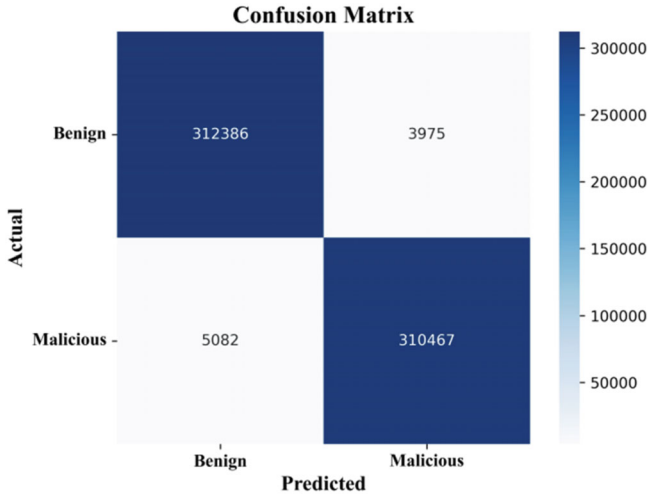


FIGURE 13 Confusion matrix of proposed FLOLSTM on Day 4.

negatives indicating an accurate identification of benign data. FLOLSTM also identified 310,467 malicious instances (True Negatives), demonstrating its effectiveness.

As can be seen in Figure 14, we are able to see the ROC curve for our proposed method in Day 4 of the analysis. The ROC curve shows that the method has high accuracy in predicting the Day 5 outcomes. This suggests that our method is effective in predicting the outcomes of similar scenarios.

4.2 | Metric-based evaluation: Proposed FLOLSTM versus existing methods

In this section, the performance of our proposed FLOLSTM is evaluated against Fuzzy-LSTM, LSTM and ANFIS in terms of accuracy, precision, recall, F-measure and specificity, FPR, and FNR. Figure 15 provides a comparative analysis of the proposed and existing techniques.

When analyzing Figure 15, our proposed "FLOLSTM" model achieves the highest accuracy of approximately 98.19%. Comparatively, "Fuzzy-LSTM" follows with an accuracy of about 91.9%, "LSTM" with 90.8%, and "ANFIS" with 89.3%. The proposed model outperforms all other models in accurately predicting both positive and negative instances.

Across precision, our Proposed "FLOLSTM" model attains about 98.24%, while "Fuzzy-LSTM" demonstrates 92.2%, "LSTM" with 91.2%, and "ANFIS" with 89.7%. For recall, "FLOLSTM" model exhibits approximately 98.19%, whereas "Fuzzy-LSTM" reaches around 92.1%, "LSTM" with 91.1%, and "ANFIS" at 89.5%.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

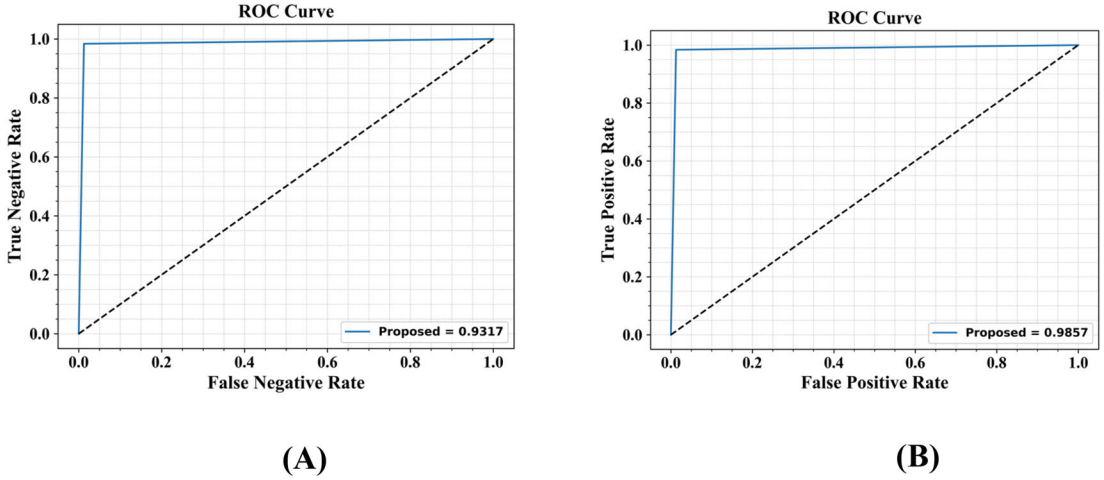


FIGURE 14 ROC curve of proposed FLOLSTM on Day 4 (A) TNR versus FNR (B) TPR versus FPR.

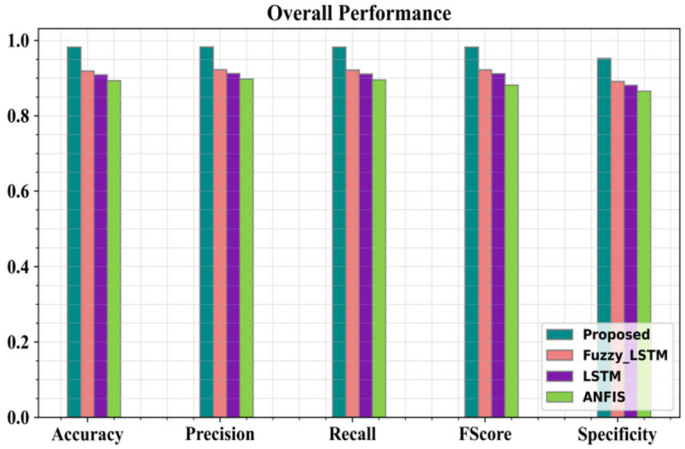


FIGURE 15 Comparative analysis of the proposed and existing techniques.

In terms of F-measure our "Proposed" model demonstrates about 98.22%, whereas "Fuzzy-LSTM" achieves around 92.2%, "LSTM" with 91.1%, and "ANFIS" at 88.1%. For specificity, our "FLOLSTM" model records approximately 95.19% specificity, while "Fuzzy-LSTM" shows around 89.1%, "LSTM" with 88.1%, and "ANFIS" at 86.5%.

The effectiveness of the "FLOLSTM" model over "Fuzzy-LSTM" might be attributed to its higher values across all metrics. This suggests that the proposed model performs better in both correctly identifying positive instances and accurately ruling out negative instances, signifying its superior predictive capability.

Based on an analysis of Figure 16, performance metrics such as FPRs and FNRs, our proposed approach is found to have a low FPR, 0.018, indicating a low probability of misclassifying benign instances as malicious.

The FNR for our proposed FLOLSTM method is 0.048, reflecting moderate misclassifications. Comparatively, the Fuzzy-LSTM method has a FPR of 0.079, suggesting higher false-positive rates, but a FNR of 0.049, which is higher rate than our method. In contrast to our proposed approach, the LSTM method displays marginally higher rates, registering 0.089 FPRs and 0.059 FNRs, suggesting a relatively high misclassification rate of benign instances.

The ANFIS method displays the highest FPR, 0.105, suggesting a pronounced tendency to misclassify benign instances as malicious, accompanied by 0.105 FNR. This shows that the proposed methodology can identify malicious instances more accurately than other approaches.

Figure 17 presents a comparative analysis of the proposed and existing techniques regarding both latency and memory usage. FLOLSTM stands out with significantly lower latency compared to other models like Fuzzy-LSTM, LSTM, and ANFIS. This decreased latency results in quicker processing times, enabling FLOLSTM to deliver rapid predictions or responses crucial for time-sensitive applications. Additionally, FLOLSTM demonstrates superior memory efficiency in comparison to Fuzzy-LSTM, LSTM, and ANFIS. Despite its lower memory requirements, FLOLSTM maintains high predictive accuracy, highlighting its ability to efficiently utilize computational resources without compromising performance.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

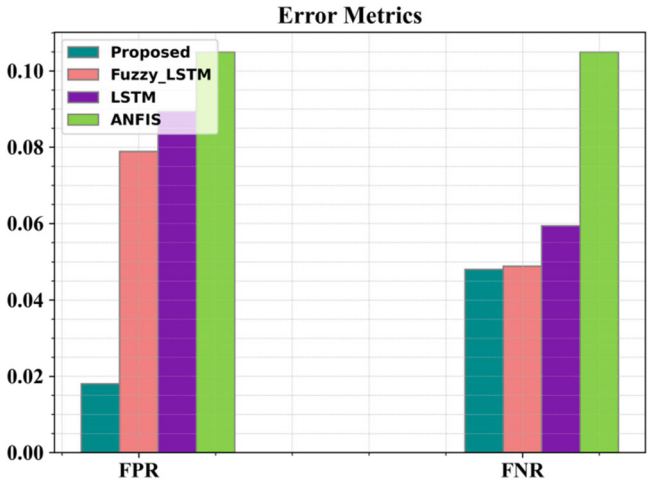


FIGURE 16 Comparative analysis of the proposed and existing techniques in terms of error metrics.

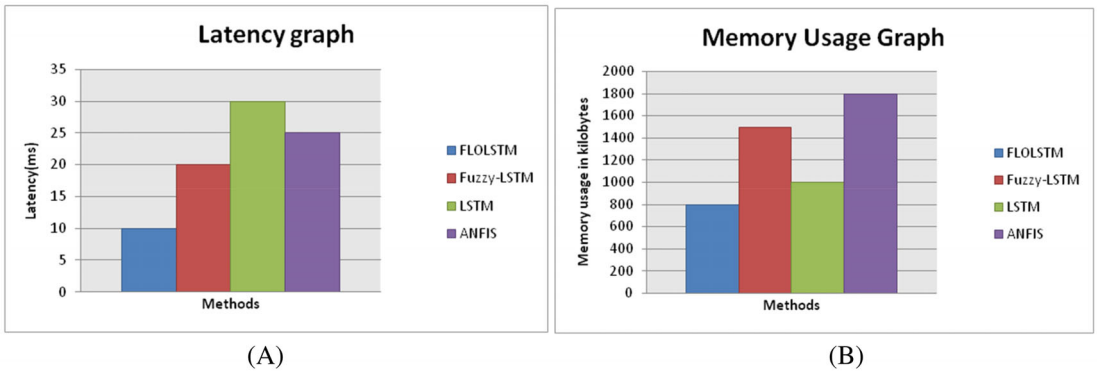


FIGURE 17 Comparative analysis of the proposed and existing techniques (A) Latency (B) Memory usage.

Therefore, the combination of low latency and minimal memory usage positions FLOLSTM as a model with high computational efficiency, making it an ideal choice for applications prioritizing speed, accuracy, and resource conservation.

The Figure 18 depicts a graph for accuracy versus epoch, while the Figure 19 depicts a graph for loss versus epoch. The graph for accuracy shows a steady increase in accuracy as the epochs increase. The graph for loss shows a decrease in loss as the epochs increase, indicating that the model is learning.

4.3 | Comparative analysis of proposed model with existing models

The comparative analysis presented in Table 2 showcases the performance metrics of the proposed FLOLSTM model in contrast to existing models for malicious traffic detection in hypervisor environments. The table provides insights into the accuracy, precision, recall, F-score, and specificity of various models, including PSI-NetVisor, WLI-FCM, FCM-SVM, IMFL-IDSCS, HIDM, and the proposed FLOLSTM model.

Among the existing models, WLI-FCM and FCM-SVM demonstrate high accuracy rates of 97.71% and 97.37%, respectively. While FCM-SVM exhibits notable precision and recall values, other models like PSI-NetVisor, CNN-GRU, and HIDM also show competitive performance in specific metrics. However, the proposed FLOLSTM model outperforms all existing models with an impressive accuracy of 98.19%, precision of 98.24%, recall of 98.19%, and F-score of 98.22%. This superior performance of the FLOLSTM model highlights its effectiveness in detecting malicious activities within hypervisor environments, showcasing its potential as a robust and reliable solution for enhancing network security. The results clearly indicate that the proposed FLOLSTM model excels in multiple key metrics compared to existing models, emphasizing its superiority in accurately identifying malicious traffic and distinguishing between benign and malicious instances.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

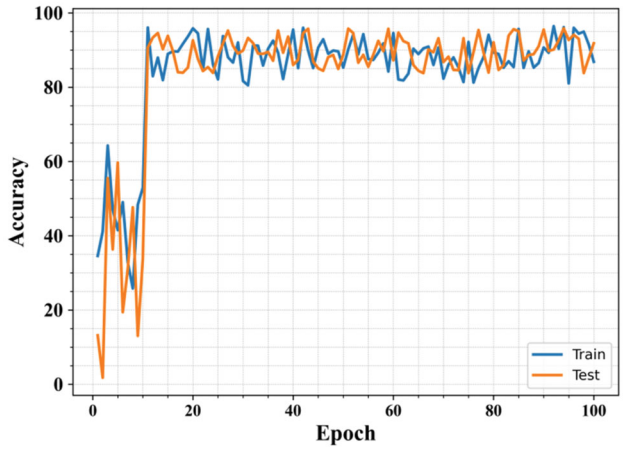


FIGURE 18 Accuracy versus Epoch graph.

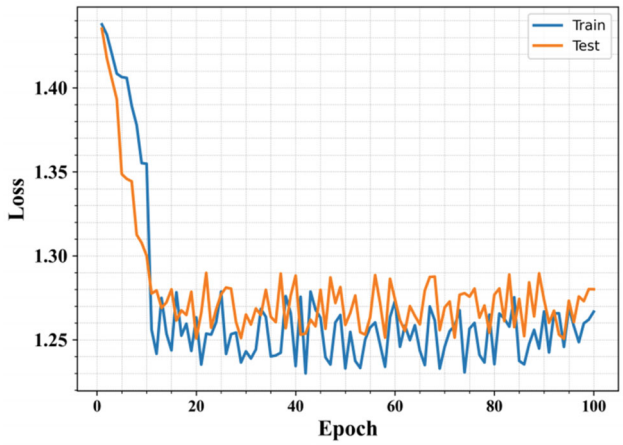


FIGURE 19 Loss versus Epoch graph.

TABLE 2 Comparative analysis of proposed model with existing models.

Existing models	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)	Specificity (%)
PSI-NetVisor[13]	94.54	-	-	-	-
WLI-FCM[18]	97.71	-	-	-	-
FCM-SVM[19]	97.37	96.20	97.90	97.00	-
IMFL-IDSCS[21]	-	92.03	78.25	81.80	-
HIDM[22]	92.7	95.1	92.7	93.5	-
SMOTE-SVM[23]	95.22	98.12	95.42	97.09	-
CNN-GRU[24]	98.16	-	97.56	-	-
SVM[25]	95.24	-	97.91	-	-
FLOLSTM [proposed]	98.19	98.24	98.19	98.22	95.19

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.4 | Discussion

During the implementation of the FLOLSTM model, several challenges emerged, each demanding unique solutions for successful deployment. One of the primary challenges faced was ensuring the quality of the data collected from hypervisor monitors. Incomplete or inaccurate data could significantly impact the effectiveness of the classifier. To address this challenge, improved data collection methods were implemented, along with rigorous data cleaning and preprocessing techniques to enhance data reliability. The FLOLSTM classifier exhibits promise in various applications, yet it is not without its limitations.

First, its performance heavily relies on the quality of data collected and preprocessed from hypervisor monitors. Incomplete or inaccurate data can significantly undermine the classifier's effectiveness, highlighting the need for improved data collection, preprocessing, and quality assurance techniques. Second, the computational complexity of the WPO algorithm poses a challenge, leading to longer processing times. Addressing these limitations presents avenues for future research. Enhancing data quality could involve refining data collection methods, implementing robust preprocessing techniques, and exploring data augmentation strategies. Similarly, optimizing the computational complexity of the optimization algorithm may require algorithmic enhancements, parallel computing approaches, or alternative optimization techniques. By tackling these challenges, future work can enhance the reliability, efficiency, and applicability of the FLOLSTM classifier across diverse domains, including real-time monitoring, anomaly detection, and predictive maintenance.

5 | CONCLUSION

In this work, we introduced Fuzzy Logic-driven Optimized Long Short-Term Memory called FLOLSTM for hypervisor-based environments. In order to enhance data reliability, data from hypervisor monitors was first carefully collected, cleaned, and preprocessed in order to improve their reliability. Following that, the preprocessed data have been sent to a FLOLSTM classifier, which will identify any malicious activities in the data that have been detected. Further, the performance of the LSTM was optimized by using a WPO algorithm known as WPO. We have compared the performance of FLOLSTM against other existing techniques such as Fuzzy-LSTM, LSTM, and ANFIS in terms of accuracy, precision, recall, F-measure, specificity, and FPR. According to the experimental results, FLOLSTM is more accurate than existing techniques, has a greater precision of 98.24%, a greater recall of 98.19%, a greater F-measure of 98.22%, and a greater specificity of 95.19%. Further, it was shown that the proposed FLOLSTM is essentially a promising device for detecting malicious activities that can be applied across a wide range of sectors. For future work, the focus could be on improving data quality through enhanced collection and efficient preprocessing methods, along with exploring optimization techniques to reduce the computational complexity of the WPO algorithm. These efforts aim to enhance the reliability and efficiency of the FLOLSTM classifier for various applications.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Anumukonda Naga Seshu Kumar https://orcid.org/0009-0006-0282-6297

REFERENCES

1. Mell P, Grance T. The NIST Definition of Cloud Computing. 2011.
2. Heady R, Luger G, Maccabe A, Servilla M. The Architecture of a Network Level Intrusion Detection System (No. LA-SUB-93-219). 1990.
3. Bhattacharya S, Maddikunta PKR, Kaluri R, et al. A novel PCA-firefly based XGBoost classification model for intrusion detection in networks using GPU. *Electronics*. 2020;9(2):219.
4. Maddikunta PKR, Parimala M, Koppu S, Gadekallu TR, Chowdhary CL, Alazab M. An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture. *Comput Commun*. 2020;160:139-149.
5. Bahram S, Jiang X, Wang Z, et al. Dksm: Subverting virtual machine introspection for fun and profit. In *2010 29th IEEE Symposium on Reliable Distributed Systems* (pp. 82-91). IEEE. 2010.
6. Chung CJ, Khatkar P, Xing T, Lee J, Huang D. NICE: network intrusion detection and countermeasure selection in virtual network systems. *IEEE Trans Dependable Secure Comput*. 2013;10(4):198-211.
7. Azab AM, Ning P, Wang Z, Jiang X, Zhang X, Skalsky NC. Hypersentry: enabling stealthy in-context measurement of hypervisor integrity. In *Proceedings of the 17th ACM Conference on Computer and communications security* (pp. 38-49). 2010.
8. Vollmar W, Harris T, Long L, Green R. Hypervisor security in cloud computing systems. *ACM Comput Surv*. 2014;1-22.
9. Scarfone K, Mell P. *Guide to Intrusion Detection and Prevention Systems (IDPS)*. Vol 800. NIST Special Publication; 2007:94.
10. Moustafa N, Hu J, Slay J. A holistic review of network anomaly detection systems: a comprehensive survey. *J Netw Comput Appl*. 2019;128:33-55.
11. Mahdavifar S, Ghorbani AA. Application of deep learning to cybersecurity: a survey. *Neurocomputing*. 2019;347:149-176.
12. Wu W, Li R, Xie G, et al. A survey of intrusion detection for in-vehicle networks. *IEEE Trans Intell Transp Syst*. 2019;21(3):919-933.
13. Mishra P, Pilli ES, Varadharajan V, Tupakula U. PSI-NetVisor: program semantic aware intrusion detection at network and hypervisor layer in cloud. *J Intell Fuzzy Syst*. 2017;32(4):2909-2921.

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14. Kotecha K, Verma R, Rao PV, et al. Enhanced network intrusion detection system. *Sensors*. 2021;21(23):7835.
15. Singh P, Ranga V. Attack and intrusion detection in cloud computing using an ensemble learning approach. *Int J Inf Technol*. 2021;13:565-571.
16. Mehmood M, Javed T, Nebhen J, et al. A hybrid approach for network intrusion detection. *CMC-Comput Mater Contin*. 2022;70:91-107.
17. Niazi RA, Faheem Y. A Bayesian game-theoretic intrusion detection system for hypervisor-based software defined networks in smart grids. *IEEE Access*. 2019;7:88656-88672.
18. Sharma P, Sengupta J, Suri PK. WLI-FCM and artificial neural network based cloud intrusion detection system. *Int J Adv Netw Appl*. 2018;10(1):3698-3703.
19. Jaber AN, Rehman SU. FCM–SVM based intrusion detection system for cloud computing environment. *Clust Comput*. 2020;23:3221-3231.
20. Rawashdeh A, Alkasassbeh M, Al-Hawawreh M. An anomaly-based approach for DDoS attack detection in cloud environment. *Int J Comput Appl Technol*. 2018;57(4):312-324.
21. Alohali MA, Elsadig M, Al-Wesabi FN, Al Duhayyim M, Mustafa Hilal A, Motwakel A. Enhanced chimp optimization-based feature selection with fuzzy logic-based intrusion detection system in cloud environment. *Appl Sci*. 2023;13(4):2580.
22. Vashishtha LK, Singh AP, Chatterjee K. HIDM: a hybrid intrusion detection model for cloud based systems. *Wirel Pers Commun*. 2023;128(4):2637-2666.
23. Wankhade N, Khandare A. Optimization of deep generative intrusion detection system for cloud computing: challenges and scope for improvements. *EAI Endorsed Trans Scalable Inf Syst*. 2023;10(6):1-13.
24. Elubeyd H, Yiltas-Kaplan D. Hybrid deep learning approach for automatic dos/DDoS attacks detection in software-defined networks. *Appl Sci*. 2023;13(6):3828.
25. Vigoya L, Pardal A, Fernandez D, Carneiro V. Application of machine learning algorithms for the validation of a new CoAP-IoT anomaly detection dataset. *Appl Sci*. 2023;13(7):4482.

How to cite this article: Kumar ANS, Yadav RK, Raghava NS. FLOLSTM: Fuzzy logic-driven optimized LSTM for improved malicious traffic detection in hypervisor environments. *Concurrency Computat Pract Exper*. 2024;e8194. doi: 10.1002/cpe.8194

15320634, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/cpe.8194 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS) | 979-8-3503-5026-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/PAIS62114.2024.10541196

Harris Hawk Optimized CLAHE and Novel Score Level Fusion of Lightweight CNNs for Multimodal Biometric Recognition

Indu Singh[1], Tanvi Singh[2], Sunaina Luthra[3], and Shruti Khatri[4]

[1,2,3,4]Department of Computer Science and Engineering, Delhi Technological University

Email: [1]indusingh@dtu.ac.in, [2]tanvisingh007official@gmail.com
[3]sunainaluthra2020@gmail.com, [4]shrutikhatri0312@gmail.com

***Abstract*—Biometric authentication is critical for secure access in diverse domains, emphasizing the need for high accuracy to prevent security breaches. Multimodal biometrics fuse traits to fortify security, vital across diverse domains, surpassing unimodal systems' vulnerability. This paper introduces a novel approach that combines Contrast Limited Adaptive Histogram Equalization (CLAHE) and Harris Hawk Optimization(HHO) (CLAHE-HHO) for image preprocessing. It is followed by an innovative score-level fusion designed for the classification of Iris and Palmprint Biometric Modalities, using Transfer Learning based Lightweight ShuffleNet and MobileNetV3 Architecture. Initially, the Iris and Palmprint images undergo preprocessing using CLAHE, with the clip limit optimized through Harris Hawk Optimization. After preprocessing, Iris images are classified using ShuffleNet, while Palmprint images undergo classification using MobileNetV3. Scores from Palmprint and Iris modalities are fused using Weighted sum Score-level Fusion, with fusion weights determined via the novel Cauchy Search-based Golden Ratio Optimization (CGRO) technique. The performance of the system is evaluated on accuracy, Detection Error Tradeoff (DET) Curve, Equal Error Rate (EER), Total Training time. The proposed methodology proves that the multimodal authentication outperformed unimodal identification with a recognition accuracy of 100 % surpassing results obtained from individual iris and palmprint modalities.**

***Index Terms*—Multimodal biometric authentication, CLAHE, Metaheuristic Optimization, Transfer Learning, Score-level Fusion**

I. Introduction

In today's digital era, traditional authentication methods like passwords and tokens are increasingly vulnerable to cyber threats. Biometric recognition offers a robust alternative, leveraging unique human characteristics for identification. Biometrics can be broadly categorized into physiological and behavioral approaches [1]. Physiological modalities (e.g., fingerprints, iris scans, and facial geometry) analyze unique physical attributes, while behavioral modalities (e.g., signature recognition, keystroke dynamics, and gait analysis) focus on individual behavioral patterns. Biometric systems authenticate individuals by capturing and analyzing their unique biometric data, comparing it against stored templates for identity verification[2]. However, relying solely on a single biometric modality (i.e. unimodal biometrics) can introduce vulnerabilities. Factors such as sensor noise, environmental variations, and potential spoofing attempts can compromise system accuracy [3].

Multimodal biometric systems address limitations by combining two or more modalities, harnessing each trait's unique strengths for enhanced authentication robustness and security [4]. For example, the combination of palmprint and iris recognition significantly improves accuracy and security compared to individual modalities [5]. Deep learning algorithms play a crucial role in extracting complex patterns from multimodal biometric data, facilitating continuous learning and adaptation to enhance precision and robustness [6].

This paper explores a multimodal biometric recognition system utilizing palmprint and iris modalities. Our approach leverages the strengths of deep learning and metaheuristics for accurate classification and identification. Firstly, Harris Hawk Optimization (HHO) fine-tunes the clip limit parameter of Contrast Limited Adaptive Histogram Equalization (CLAHE) for image contrast enhancement, improving the quality of iris and palmprint images. Subsequently, efficient transfer learning-based ShuffleNet and MobileNetV3 architectures are employed for classification of iris and palmprint images, respectively. Finally, a novel cauchy search-based golden ratio optimization is employed for score level fusion, combining the classification scores from the individual models to enhance robustness and accuracy. The proposed methodology is thoroughly evaluated using various performance metrics.

The key findings of this research are outlined below:

1) *A Novel combination of CLAHE optimized by Harris Hawk Optimization is proposed for image contrast enhancement.*
2) *Efficient transfer learning based ShuffleNet and MobileNetV3 Lightweight Convolutional Neural Network (CNN) architectures are used for classification.*
3) *An innovative Cauchy Search based Golden Ratio Optimized Score level fusion is used for combining the scores from palmprint and iris modalities.*

The paper is organized as follows: Section II provides an overview of previous studies. Section III outlines the proposed methodology, including preprocessing, classification, and score level fusion. Section IV discusses the experimental setup, while Section V reports the obtained results. Section VI summarizes key findings and proposes future research directions, concluding the paper.

979-8-3503-5026-5/24/$31.00 ©2024 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 09:57:22 UTC from IEEE Xplore. Restrictions apply.

II. Literature Review

Numerous studies have explored various techniques and methodologies to improve the accuracy, robustness, and efficiency of biometric recognition systems. In this section, we review the existing literature related to multimodal biometric recognition, metaheuristic algorithms, and deep learning models.

Image enhancement, particularly contrast enhancement, plays a significant role in biometric-recognition systems. Lionnie et al. [7] conducted a study on contrast enhancement methods in various color spaces for biometric recognition systems. They evaluated techniques such as mathematical morphology, histogram equalization, image adjustment and CLAHE in the HSV, YCbCr, and RGB color spaces. Their findings indicated that histogram equalization yielded the best results, achieving approximately 89% precision in recognition. Fan et al. [8] introduced a method for smart image enhancement utilizing CLAHE based on an F-Shift transformation during decompression.

Many researchers have used CLAHE with predefined parameters for biometric recognition (Rahman et al.[9]; Ganesan et al.[10]; Devi et al.[11]). Automatic parameter selection for CLAHE is an area of limited exploration in the literature. However, Surya and Muthukumaravel [12] introduced Adaptive Sailfish Optimization-Contrast Limited Adaptive Histogram Equalization (ASFO-CLAHE) for hyperparameter tuning in image enhancement. This study focuses on improving the mammography image quality for early breast cancer detection.

Chollet [13] emphasized that deep learning models are often not trained from scratch due to the substantial datasets and time-intensive procedures. Researchers typically prefer pre-trained models like Inception, VGGNet, AlexNet, and DenseNet to expedite training. Training from scratch entails complex experimentation with parameters like weights, filters, and layers, which can be impractical for many applications. Nguyen et al. [14] demonstrated the effectiveness of utilizing five distinct pre-trained CNN models for iris recognition. The selected models included VGGNet, Inception, AlexNet, ResNet, and DenseNet, and their performances were compared.

Michele et al. [15] investigated the utilization of MobileNet V2 deep convolutional neural networks for palmprint recognition. Specifically, they explored the effectiveness of fine-tuning a pre-trained MobileNet neural network to adapt it to palmprint recognition tasks. Zang et al. [16] introduced a novel approach called Shuffle Dense Networks (SDN) for biometric authentication, combining ShuffleNet's computational efficiency and DenseNet's hierarchical feature utilization to enhance system performance.

In multimodal biometrics, integrating data from various modalities is crucial. Lip and Ramli [17] outline three fusion levels: feature, score, and decision. Feature-level fusion concatenates modalities' features before pattern matching, while score-level fusion averages biometric scores after pattern matching. Decision-level fusion combines modalities' final decisions using logic AND and OR operations. The study highlights the effectiveness of score-level fusion in discriminating between genuine and imposter scores compared to other fusion levels.

Peng et al. [18] simplified multimodal biometric score fusion by utilizing t-norms, reducing computational complexity while enhancing authentication accuracy. Eskandari and Toygar [19] improved recognition accuracy by combining face and iris modalities, employing particle swarm optimization (PSO) for feature selection and the weighted-sum rule for score fusion. Anzar and Sathidevi [20] proposed an optimization scheme for determining integration weights in multibiometric systems with score-level fusion, optimizing weight factors using linear programming techniques to improve recognition accuracy with fingerprint and voice biometrics.

III. Proposed Methodology

The proposed methodology introduces a novel combination leveraging Harris Hawk Optimization (HHO) to optimize the clip limit parameter of Contrast Limited Adaptive Histogram Equalization (CLAHE) for the preprocessing of Iris and Palmprint datasets. Following the preprocessing, Iris images undergo classification using ShuffleNet and Palmprint images are classified using MobileNetV3. The classification scores from both models are ensembled using a novel Cauchy based Golden Ratio Optimized (C-GRO) Score level fusion. The architecture of the proposed model is depicted in figure 1.

A. Image Preprocessing

The proposed model implements the novel combination of CLAHE and Harris Hawk Optimization (HHO). CLAHE [21] is a histogram equalization technique applied locally to enhance the contrast of images while limiting the amplification of noise in specific regions. Clip limit parameter sets the threshold above which histogram bins are uniformly redistributed, preventing excessive contrast enhancement.

Harris Hawk Optimization [22] is a relatively new population based metaheuristic algorithm. It is a nature-inspired optimization algorithm that mimics the collaborative hunting behavior of Harris's hawks.

The objective function measures the Features Similarity Index Matrix (FSIM) between the original and CLAHE-enhanced images. The FSIM includes calculations for the Gradient Magnitude (GM) features and Phase Congruency (PC). PC evaluates the similarity of local patterns and edge information using corner harris[23] while GM is a measure of the strength of intensity variations in the images based on Scharr operator.

$$FSIM(x,y)= \frac{2\cdot T_1\cdot PC_a(x,y)+T_2\cdot GM_a(x,y)}{T_1\cdot PC_a(x,x)+T_2\cdot GM_a(x,x)+T_1\cdot PC_a(y,y)+T_2\cdot GM_a(y,y)+\varepsilon} \quad (1)$$

$$PC_a(x,y) = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i x_i^2}\cdot\sqrt{\sum_i y_i^2}+\varepsilon} \quad (2)$$

$$GM_a(x,y) = \frac{2\cdot\sum_i \sqrt{x_i^2+y_i^2}}{\sum_i \sqrt{x_i^2}+\sum_i \sqrt{y_i^2}+\varepsilon} \quad (3)$$

where x and y represent corresponding pixels in the two images being compared. T_1 and T_2 are positive constants

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 09:57:22 UTC from IEEE Xplore. Restrictions apply.

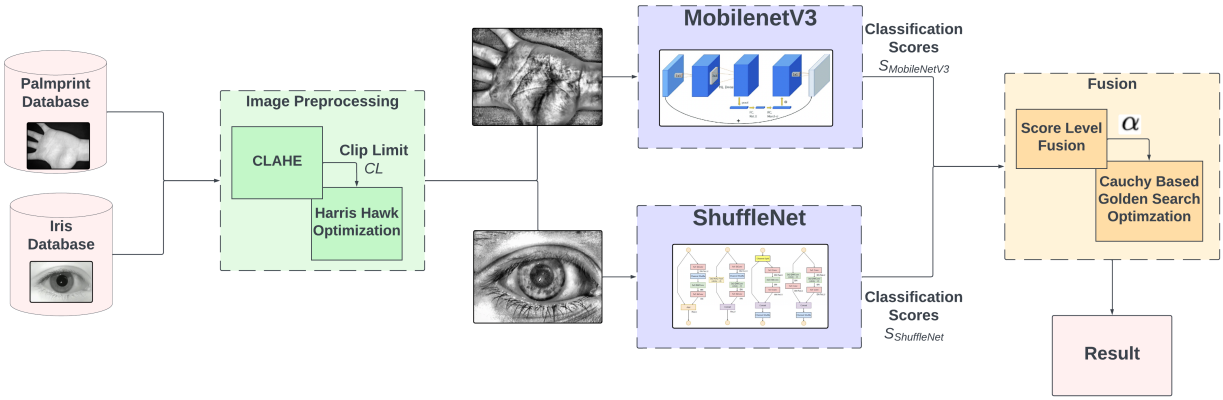


Fig. 1: Proposed Model

controlling the relative importance of PC and GM, typically set to 1. $PC_a(x, y)$ and $GM_a(x, y)$ denote the normalized similarity measures for PC and GM, respectively. ϵ is a small constant to avoid division by zero. FSIM values range from 0 to 1, with higher values indicating greater similarity between the images.

Hunting behavior of Harris hawk can be divided into 2 phases: *Exploration* and *Exploitation*. HHO dynamically balances exploration and exploitation based on a parameter called "Levy Flight or Espace Energy(E)".

$$E = |F_{rabbit(t)} - F_{avg(t)}| \tag{4}$$

Where, $F_{rabbit(t)}$ is the fitness of the rabbit at current iteration and $F_{avg(t)}$ is the average fitness of hawks. The rabbit is the leader hawk in the population.

1) **Exploration**
When the energy is high, exploration is prioritized. Hawks use a Levy flight mechanism to search for prey in a wider area. Levy flights allow for long jumps, enabling exploration of distant regions in the search space. The updated position of each hawk is calculated as:

$$X_i(0) = \text{rand}(1, d) \cdot (ub - lb) + lb \tag{5}$$

$$X_i(t{+}1) = X_{rabbit(t)} - S{*}Levy(d){*}(X_{rabbit(t)} - X_i(t)) \tag{6}$$

$$Levy(x) = 0.01 \times \frac{u \times \sigma}{|v|^{\frac{1}{\beta}}} \tag{7}$$

$$\sigma = \left(\frac{\Gamma(1+\beta) \times \sin(\frac{\pi\beta}{2})}{\Gamma(\frac{1+\beta}{2}) \times \beta \times 2^{(\frac{\beta-1}{2})}}\right)^{\frac{1}{\beta}} \tag{8}$$

Where, ub and lb are the upper and lower bounds respectively, $X_i(t+1)$ is the position vector of hawks in the next iteration t, $X_{rabbit(t)}$ is the position of rabbit, $X_i(t)$ is the current position vector of hawks, S is a random vector by size $1 \times d$, $Levy(d)$ is the levy flight function with d as the dimension of the problem, u, v are random values inside (0,1) and β is a default constant set to 1.5.

2) **Exploitation**
In the exploitation phase, hawks focus on refining their positions around the current best solution. Two main strategies are employed:
a) Soft Besiege: The hawks gradually encircle the prey, progressively reducing the search space. The updated position is calculated as:

$$X_{i(t+1)} = X_{rabbit(t)} - |Q| * E * (X_{rabbit(t)} - X_{i(t)}) \tag{9}$$

Where, Q is a random vector uniformly distributed between [-1, 1].
When the Levy flight energy value is low but not zero, the algorithm chooses to apply the "rapid dives" strategy. It involves attempting two jumps with varying degrees of randomness and modifying the hawk's position based on these jumps, while ensuring the positions stay within the specified boundaries.
b) Hard Besiege: If the prey escapes, a sudden attack (hard besiege) is simulated by randomly choosing a hawk from the population (excluding the best one) and updating the prey's position:

$$X_{rabbit(t+1)} = X_{rand(t)} - R{*}abs(D_{rabbit(t)}){*}levy(d) \tag{10}$$

Where, $X_{rand(t)}$ is the position of a randomly chosen hawk, R is a random scaling factor uniformly distributed between [0, 2] and $D_{rabbit(t)}$ is the difference between the current prey position and the average position of all hawks.

Algorithm 1 demonstrates working of the harris hawk optimization algorithm. The search space for clip limit ranges between [1, 40] , and the number of hawks is set to 50 while total number of iterations for finding the optimal clip limit equals 10. Figure 2 and 3 shows the original and preprocessed images. These image grey level features are further used for classification.

B. Classification

The proposed architecture employs two different lightweight neural architecture for classification of Palmprint and Iris biometrics to provide a contrast between the efficiencies of the neural networks and leverage advantages of both to provide an accurate and computationally efficient model. The preprocessed palmprint dataset undergoes classification through mobileNetV3[24] while iris

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 09:57:22 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Harris Hawks Optimization

Input: $PopulationSize$, $MaxIterations$, $Dimension$

Output: Best solution: X_{Rabbit}

Initialize $hawks$ population, with randomized values in $[0, 1]$ for each dimension;
for $t \leftarrow 1$ **to** Max_iter **do**
  for *Each* $hawk$ *in* $hawks$ **do**
    Perform Levy Flight;
    $L[hawk] \leftarrow levy(hawk)$;
    Evaluate Fitness;
    $F[hawk] = FSIM(hawk)$;
  $X_{rabbit} \leftarrow$ best location of rabbit;
  for *Each* $hawk$ *in* $hawks$ **do**
    Update $E \leftarrow Eq4$;
    if $|E| \geq 1$ **then**
      Exploration Phase;
      Update $X_{hawk} \leftarrow Eq(6)$;
    if $|E| < 1$ **then**
      Generate random probability P;
      if $P \geq 0.5$ *and* $|E| \geq 0.5$ **then**
        Soft Besiege;
        Update $X_{hawk} \leftarrow Eq(8)$;
      else if $P \geq 0.5$ *and* $|E| < 0.5$ **then**
        Hard Besiege;
        Update $X_{hawk} \leftarrow Eq(9)$;
      else if $P < 0.5$ *and* $|E| \geq 0.5$ **then**
        Soft Besiege With rapid dives;
      else if $P < 0.5$ *and* $|E| < 0.5$ **then**
        Hard Besiege With rapid dives;
  Limit solutions within boundaries for each $hawk$ in $hawks$;
return X_{Rabbit};

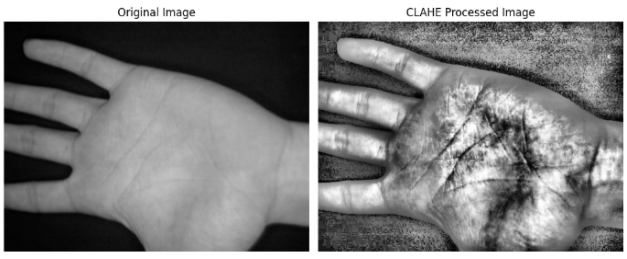


Fig. 2: Original and Preprocessed Palmprint Image

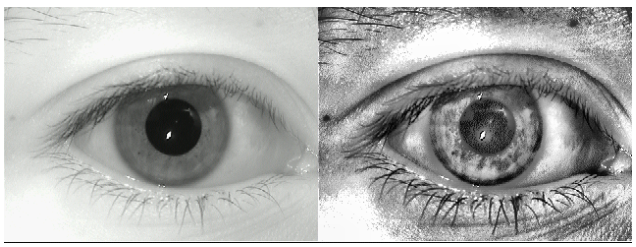


Fig. 3: Original and Preprocessed Iris Image

dataset uses shuffleNet[25] for classification. Both ShuffleNet and mobileNetV3 are lightweight CNN architecture suitable for mobile and edge devices.

MobileNetV3 prioritizes efficient feature extraction and leverages the Squeeze-and-Excitation (SE) blocks for adaptive recalibration of channel-wise feature representations. Mathematical equation for an SE block:

$$SE(x) = \sigma(F(pool(x))) * x \tag{11}$$

where,
x: Input feature map
$pool$: Global average pooling
F: Fully-connected layer with reduction ratio
σ: Sigmoid activation.

These blocks enhance the model's capacity to emphasize informative features. The architecture adopts the H-Swish activation function, a blend of ReLU and hard-tanh, offering improved performance with reduced computational cost.

$$\text{h-swish}\,(x) = x\frac{\text{ReLU6}\,(x+3)}{6} \tag{12}$$

It incorporates efficient inverted bottleneck blocks, employing depthwise separable convolutions and linear bottlenecks to reduce computational complexity while preserving accuracy.

Other classification technique used is shuffleNet. It optimizes computational efficiency by utilizing group convolutions and channel shuffling. Channel shuffling involves rearranging feature maps along the channel dimension, promoting efficient communication across the network while preserving performance. The channel shuffle operation is mathematically defined as follows:

$$\text{Step 1: } Y = \text{reshape}(X, (B, G, \frac{C}{G}, H, W)) \tag{13}$$

$$\text{Step 2: } Z = \text{transpose}(Y, (0, 2, 1, 3, 4)) \tag{14}$$

$$\text{Step 3: Shuffle}(X) = \text{reshape}(Z, (B, C, H, W)) \tag{15}$$

Here, B is the batch size, C is the number of channels, H is the height, W is the width, and G is the number of groups. Apart from the channel shuffle operation, ShuffleNet also incorporates depth-wise separable convolutions and point-wise group convolutions to further reduce computation and parameters.

C. CGRO Score Level Fusion

A score-level fusion approach is employed where the classification scores obtained from each network are combined to make a final decision. The fused score S_{fusion} can be computed as:

$$S_{fusion} = \alpha * S_{MobileNetV3} + (1-\alpha) * S_{shuffleNet} \tag{16}$$

where α are randomly assigned weights fine tuned using novel Cauchy based Golden Ratio Optimization for robust classification.

The golden ratio optimization exploits the mathematical constant known as the golden ratio (φ).

$$\varphi = (1 + \sqrt{5})/2 \approx 1.61803398875 \tag{17}$$

After intializing golden ratio left and right exploration is prioritized.

$$c = b - \varphi(b - a) \tag{18}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 09:57:22 UTC from IEEE Xplore. Restrictions apply.

$$d = a + \varphi(b - a) \tag{19}$$

$$\alpha = \frac{c + d}{2} \tag{20}$$

Where, a, b are exploration interval bounds; c, d are Golden Ratio derived exploration bounds within $[a, b]$; α is the midpoint of exploration interval.

CGRO exploits cauchy distribution to generate step size for exploration in GRO. Cauchy distribution defines the step size as :

$$step_size = \gamma \cdot \tan(\pi \cdot (u - 0.5)) \tag{21}$$

where u is a random number between 0 and 1, γ is the scale parameter which influences the spread or width of the distribution and lies in $[0, \infty]$. We optimize score-level fusion with CGRO to find the optimal α value using accuracy as the objective function, and exploit the fused score to calculate final results.

IV. Experimental Setup

The proposed methodology was executed in Python3 on Google Colab CPU runtime with 12.7 GB system RAM. A single class represents an individual and has been trained on dataset comprising 10 images, including both right and left images. The database for iris was obtained from MMU1[26] which consists of 5 images each of left and right IRIS of 46 subjects, totalling 460 images. CASIA-palmprint V1[27] is used as a dataset for palmprint images. It contains 5,502 images captured from 312 subjects. 8 left palm images and 8 right palm images were obtained from each subject. In this work, a novel multimodal database MMU-CAS has been created by combining iris images of 45 subjects selected from the MMU1 with palmprint images of another 45 subjects selected from the CASIA-palmprint V1. Another palm-print dataset-IIT Delhi Touchless palmprint [28], consists of left and right hand images of 230 subjects. Hence, a multimodal database MMU-IIT has been created using the same method used for preparation of MMU-CAS database. 360 iris images from MMU-iris database were trained using pre-trained-shuffleNet CNN architecture, while 90 images of both left and right iris were used for testing and evaluating performance metrics. Similarly, 450, 360 palmprint images from CASIA, IIT respectively, were trained using pre-trained MobileNetV3 neural network architecture and performance was evaluated on 90 test images each. The existing model will be utilized to train new individuals without retraining the entire model. The Train to test ratio that has been maintained for every biometric dataset is 80:20.

V. Result

The performance of the proposed methodology is evaluated in terms of Accuracy, EER, FRR, FAR, Macro F1 score and total training time.

A. Impact of preprocessing on performance Metrics

Table I illustrates the impact of preprocessing on the performance metrics of the model. It is evident that a preprocessed biometric database yields superior performance concerning accuracy, FAR, FRR, and F1 score, all achieved within a relatively shorter training time as compared to unprocessed dataset.

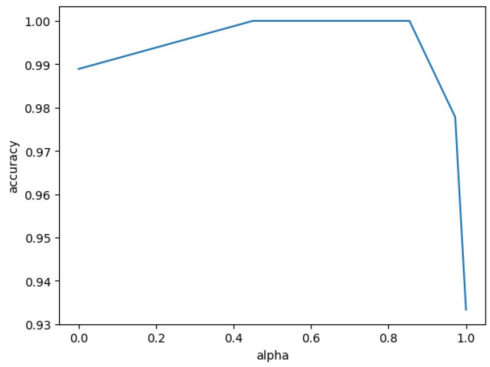
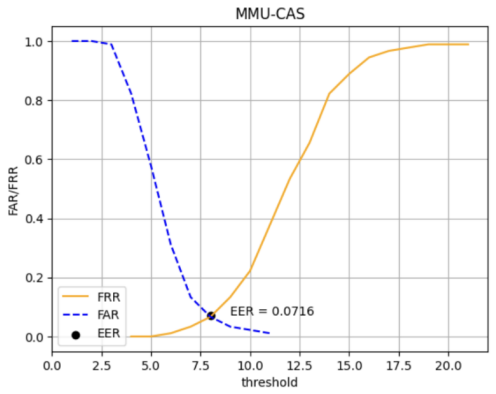
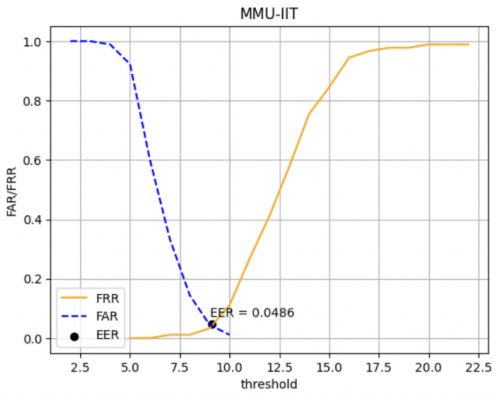


Fig. 4: Accuracy vs Score-level Parameter



(a) MMU-CASIA



(b) MMU-IIT

Fig. 5: EER Graphs

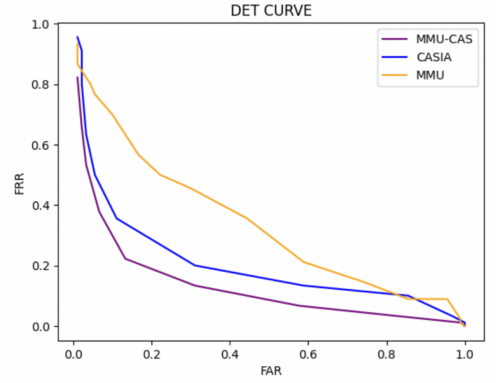


Fig. 6: DET Curve

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 09:57:22 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Impact of Preprocessing

Dataset	Preprocessing	Cliplimit	ACC	FRR	FAR	Macro F1 Score	Total Training Time
MMU	Yes	20.62	98.88%	0.0111 ± 0.0737	0.00025 ± 0.00167	0.98814	6 min 5 sec
MMU	No	-	86.84%	0.0018	0.1306	0.8177	15 min 9 sec
CASIA	Yes	40.211	93.33%	0.0666 ± 0.1699	0.001515 ± 0.0033	0.9288	20 min 51 sec
CASIA	No	-	84.44%	0.02	0.189	0.8177	25 min 9 sec
IIT	Yes	15.211	95.55%	0.0212 ± 0.0818	0.00034 ± 0.00156	0.9592	10 min 9 sec
IIT	No	-	86.44%	0.02 ± 0.098	0.0014 ± 0.016	0.8587	12 min 17 sec

TABLE II: Results of Different CNN Models

Model	ACC	Train Loss	Val Loss	Total Training Time	Training Time per Epoch	Input Size	Total Num of Features
InceptionV3	77.81%	0.0001	0.0154	38 min	5.6 min	(3,299,299)	2,68,203
MobileNetV3	96.05%	0.0001	0.0034	12 min	50 sec	(3,224,224)	1,50,528
ShuffleNet	98.88%	0.0001	0.0025	6 min 5 sec	31 sec	(3,224,224)	1,50,528
Densenet121	100%	0.0001	0.0006	25 min 33 sec	2.14 min	(3,224,224)	1,50,528
Resnet50	97.64%	0.0002	0.0020	27.4 min	2.5 min	(3,224,224)	1,50,528

TABLE III: Impact of Transfer Learning

Dataset	Model	Training Size	Transfer Learning	ACC	FRR	FAR
MMU	ShuffleNet	(360,224,224)	Yes	98.88%	0.0111 ± 0.0737	0.00025 ± 0.00167
MMU	ShuffleNet	(360,224,224)	No	33.33%	0.666 ± 0.3801	0.01515 ± 0.02419
IIT	MobileNetV3	(360,224,224)	Yes	95.55%	0.0212 ± 0.0818	0.00034 ± 0.00156
IIT	MobileNetV3	(360,224,224)	No	2.5%	0.9778 ± 0.103	0.0222 ± 0.060
CASIA	MobileNetV3	(450,224,224)	Yes	93.33%	0.0666 ± 0.1699	0.001515 ± 0.0033
CASIA	MobileNetV3	(450,224,224)	No	48.88%	0.511 ± 0.4012	0.161 ± 0.021

TABLE IV: Unimodal and Multimodal Database Metrics

Databset	Accuracy	FRR	FAR	F1-Score
CASIA	93.33%	0.0666 ± 0.1699	0.001515 ± 0.0033	0.9288
MMU	98.88%	0.0111 ± 0.0737	0.00025 ± 0.00167	0.98814
IIT	95.55%	0.0212 ± 0.0818	0.00034 ± 0.00156	0.9592
MMU-CAS	100%	0	0	1
MMU-IIT	100%	0	0	1

TABLE V: Equal Error Rate (EER) and Threshold values for multiple datasets

Dataset	EER	THRESHOLD	ACCURACY
CASIA	0.44	3.9	93.34%
MMU	0.19	16.4	98.89%
IIT	0.08	2	95.56%
MMU-CAS	0.0716	8	100%
MMU-IIT	0.0486	9	100%

TABLE VI: Comparison of our approach with other past approaches

Author	Year	Technique used	Accuracy
Xu et al. [29]	2020	CNN	98.84%
Hsiao et al.[30]	2021	VGG-16	98%
Xu et al.[31]	2022	PRSCNet	99.2%
Mukahar [32]	2023	PCANet	99.77%
Alshakree et al.[33]	2023	GoogLeNet + SqueezeNet + AlexNet + GWO	96.72%
Our Approach	**2024**	**CLAHE-HHO + Lightweight CNN + CGRO**	**100%**

B. Results of Preprocessed Unimodal Biometric Recognition using Different Models

The performance metrics of multiple models trained on MMU iris dataset are summarized in Table II. Analysis of Table II reveals that Densenet achieves the highest accuracy, followed by ShuffleNet. However, it is noteworthy that training the Densenet model requires a total of 25 minutes, with an average of 2.14 minutes per epoch. This is significantly higher compared to the total training time of the ShuffleNet and MobileNetV3 models. As a result, we fused the scores predicted by the ShuffleNet model trained on iris images and the MobileNetV3 model trained on palmprint images, achieving an accuracy of 1, as demonstrated in Table IV.

Table III compares test results with and without transfer learning for various datasets. It demonstrates that transfer learning yields superior performance and stability in biometric recognition.

C. Comparison of performance of Unimodal and Multimodal Biometric datasets

Table IV illustrates the accuracy achieved for the CASIA palmprint dataset trained on MobileNetV3 (93.34%) and the MMU iris database trained on ShuffleNet (98.89%). Remarkably, the accuracy of the multimodal MMU-CAS dataset is 100 %c, outperforming the unimodal results. Similar results are observed for the MMU-IIT dataset as well.

To evaluate the performance of the multimodal MMU-CAS database, the final score used for class prediction is calculated as the average of the scores predicted by the ShuffleNet(Iris) and MobileNetV3(Palmprint) models.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 09:57:22 UTC from IEEE Xplore. Restrictions apply.

Subsequently, the class with the highest score is assigned to the respective test image. In equation 16, we set alpha to 0.5 after observing figure 4 that accuracy peaks (100%) within an alpha range of 0.45 to 0.85, with deviations resulting in decreased accuracy.

FAR (False Acceptance Rate) arises when an imposter is mistakenly accepted as genuine, while FRR (False Rejection Rate) occurs when a genuine person is wrongly rejected. EER (Equal Error Rate) indicates the point where FAR equals FRR. The decision threshold is set at this EER value. Scores below this threshold are classified as imposters, while those above are classified as genuine. Table V shows that the EER for multimodal databases is lower compared to unimodal datasets and hence perform better. The FAR/FRR vs Threshold graph (Figure 5) helps in evaluating the EER and decision threshold.

The performance of both unimodal and multimodal datasets were evaluated using the Detection Error Tradeoff(DET) curve, which is a graphical representation that illustrates the relationship between FAR and FRR. Figure 6 demonstrates that the FAR-FRR tradeoff for the MMU-CAS multimodal database is better than that of the unimodal MMU and CASIA datasets.

D. Comparison with previous studies

Table VI highlights the robustness of our model by comparing its results with those from other recent models. In 2020, Xu et al. [29] achieved an accuracy of 98.84% with an end-to-end CNN model for multimodal biometric recognition. In 2021, Hsiao et al. [30] used a YOLO-based detector and VGG-16 model for iris recognition, achieving 98% accuracy. In 2022, Xu et al. [31] introduced the PRSCNet (Palmprint Recognition Fusing Spatial and Channel Features Network) model with SE (Squeeze-and-Excitation) modules for palmprint recognition, attaining 99.2% accuracy. In 2023, Mukahar [32] used PCANet-based deep learning and classic techniques, reaching 99.77% accuracy and Alshakree et al. [33] combined GoogLeNet, SqueezeNet, and AlexNet with gray wolf optimization for palm print analysis, achieving 96.72% accuracy. Table VI shows that our proposed model, CLAHE-HHO + Lightweight CNN + CGRO, achieves the highest accuracy and surpasses previous models.

VI. CONCLUSION

For efficient and comprehensive recognition, our research employs a multimodal biometric system utilizing a novel combination of CLAHE, with the cliplimit determined by the HHO metaheuristic technique for image contrast enhacement followed by transfer learning based ShuffleNet and MobileNetV3 architectures for classification of iris and palmprint images respectively. Finally a novel CGRO is used to determine the weight for score level fusion that combines the scores from palmprint and iris modalities. Our approach demonstrates high efficacy by achieveing 100% accuracy with lower EER, FAR, and FRR compared to unimodal systems. The DET curve for our multimodal dataset outperforms individual iris or palmprint datasets. Moreover, by employing lightweight deep learning models, our system achieves superior performance in less training time compared to other CNN models, making it highly efficient. The scalability of this approach to recognize individuals depends on dataset diversity, potentially accommodating a large number of classes/individuals but subject to computational resources and task complexity.

As a future improvement, the equal error rate can further be reduced by incorporating more biometric traits like finger vein, palm vein, face features etc for biometric recognition and authentication.

REFERENCES

[1] Shaymaa Adnan Abdulrahman and Bilal Alhayani. A comprehensive survey on the biometric systems based on physiological and behavioural characteristics. *Materials Today: Proceedings*, 80:2642–2646, 2023.

[2] Alfred C Weaver. Biometric authentication. *Computer*, 39(2):96–97, 2006.

[3] Muhtahir O Oloyede and Gerhard P Hancke. Unimodal and multimodal biometric sensing systems: a review. *IEEE access*, 4:7532–7555, 2016.

[4] Kunlei Jing, Xinman Zhang, and Xuebin Xu. An overview of multimode biometric recognition technology. In *Proceedings of the 6th International Conference on Information Technology: IoT and Smart City*, pages 168–172, 2018.

[5] Yakshita Jain and Mamta Juneja. A comparative analysis of iris and palm print based unimodal and multimodal biometric systems. In *Innovations in Computer Science and Engineering: Proceedings of the Fourth ICICSE 2016*, pages 297–306. Springer, 2017.

[6] Riseul Ryu, Soonja Yeom, Soo-Hyung Kim, and David Herbert. Continuous multimodal biometric authentication schemes: a systematic review. *IEEE Access*, 9:34541–34557, 2021.

[7] Regina Lionnie, Fyan Estu Widyantoro, and Mudrik Alaydrus. A study of contrast enhancement methods in various color spaces on biometric recognition system. In *2019 Fourth International Conference on Informatics and Computing (ICIC)*, pages 1–4. IEEE, 2019.

[8] Ruiqin Fan, Xiaoyun Li, Sanghyuk Lee, Tongliang Li, and Hao Lan Zhang. Smart image enhancement using clahe based on an f-shift transformation during decompression. *Electronics*, 9(9):1374, 2020.

[9] Muhammad Fauzan Rahman, Febryanti Sthevanie, and Kurniawan Nur Ramadhani. Face recognition in low lighting conditions using fisherface method and clahe techniques. In *2020 8th International Conference on Information and Communication Technology (ICoICT)*, pages 1–6. IEEE, 2020.

[10] Thenmozhi Ganesan, Anandha Jothi Rajendran, and Palanisamy Vellaiyan. An efficient finger vein image enhancement and pattern extraction using clahe and repeated line tracking algorithm. In *Intelligent Computing, Information and Control Systems: ICICCS 2019*, pages 690–700. Springer, 2020.

[11] D Anjani Suputri Devi, D Sasi Rekha, Mudugu Kishore Kumar, P Rama Mohana Rao, and G Naga Vallika. Transfer learning-based

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 09:57:22 UTC from IEEE Xplore. Restrictions apply.

effective facial emotion recognition using contrast limited adaptive histogram equalization (clahe). In *International Conference on Computer Vision, High-Performance Computing, Smart Devices, and Networks*, pages 273–286. Springer, 2022.

[12] S Surya and A Muthukumaravel. Adaptive sailfish optimization-contrast limited adaptive histogram equalization (asfo-clahe) for hyperparameter tuning in image enhancement. In *Computational Intelligence for Clinical Diagnosis*, pages 57–76. Springer, 2023.

[13] Francois Chollet. *Deep learning with Python*. Simon and Schuster, 2021.

[14] Kien Nguyen, Clinton Fookes, Arun Ross, and Sridha Sridharan. Iris recognition with off-the-shelf cnn features: A deep learning perspective. *IEEE Access*, 6:18848–18855, 2017.

[15] Aurelia Michele, Vincent Colin, and Diaz D Santika. Mobilenet convolutional neural networks and support vector machines for palmprint recognition. *Procedia Computer Science*, 157:110–117, 2019.

[16] Huaijuan Zang, Qiuyu Li, Ailian Chen, Jinlong Chen, and Shu Zhan. Shuffle dense networks for biometric authentication. In *Cyberspace Safety and Security: 11th International Symposium, CSS 2019, Guangzhou, China, December 1–3, 2019, Proceedings, Part I 11*, pages 143–156. Springer, 2019.

[17] Chia Chin Lip and Dzati Athiar Ramli. Comparative study on feature, score and decision level fusion schemes for robust multibiometric systems. *Frontiers in computer education*, pages 941–948, 2012.

[18] Jialiang Peng, Ahmed A Abd El-Latif, Qiong Li, and Xiamu Niu. Multimodal biometric authentication based on score level fusion of finger biometrics. *Optik*, 125(23):6891–6897, 2014.

[19] Maryam Eskandari and Önsen Toygar. Score level fusion for face-iris multimodal biometric system. In *Information Sciences and Systems 2013: Proceedings of the 28th International Symposium on Computer and Information Sciences*, pages 199–208. Springer, 2013.

[20] Sharafudeen Mohammed Anzar and PS Sathidevi. Optimization of integration weights for a multibiometric system with score level fusion. In *Advances in Computing and Information Technology: Proceedings of the Second International Conference on Advances in Computing and Information Technology (ACITY) July 13-15, 2012, Chennai, India-Volume 2*, pages 833–842. Springer, 2013.

[21] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.

[22] Ali Asghar Heidari, Seyedali Mirjalili, Hossam Faris, Ibrahim Aljarah, Majdi Mafarja, and Huiling Chen. Harris hawks optimization: Algorithm and applications. *Future generation computer systems*, 97:849–872, 2019.

[23] Gugulethu Mabuza-Hocquet and Fulufhelo Nelwamondo. Fusion of phase congruency and harris algorithm for extraction of iris corner points. In *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, pages 315–320. IEEE, 2015.

[24] Brett Koonce and Brett Koonce. Mobilenetv3. *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pages 125–144, 2021.

[25] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

[26] MMU iris dataset. https://www.kaggle.com/datasets/naureenmohammad/mmu-iris-dataset.

[27] CASIA Palmprint Image Database. http://www.idealtest.org/dbDetailForUser.do?id=5#/datasetDetail/5.

[28] IIT Delhi Palmprint Image Database version 1.0. http://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Palm.htm.

[29] Hui Xu, Jun Kong, Miao Qi, and Yinghua Lu. Multimodal biometric recognition based on convolutional neural networks. In *Artificial Intelligence and Security: 6th International Conference, ICAIS 2020, Hohhot, China, July 17–20, 2020, Proceedings, Part II 6*, pages 112–123. Springer, 2020.

[30] Cheng-Shun Hsiao, Chih-Peng Fan, and Yin-Tsung Hwang. Iris location and recognition by deep-learning networks based design for biometric authorization. In *2021 IEEE 3rd Global conference on life sciences and technologies (LifeTech)*, pages 144–145. IEEE, 2021.

[31] Xuebin Xu, Meijuan An, and Yibiao Liu. A method of color palmprint recognition based on densenet integrate spatial and channel features. In *2022 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 975–980. IEEE, 2022.

[32] Nordiana Mukahar. Performance comparison of pcanet-based deep learning techniques for palmprint recognition. In *2023 IEEE International Conference on Applied Electronics and Engineering (ICAEE)*, pages 1–6. IEEE, 2023.

[33] Firas Alshakree, Ayhan Akbas, and Javad Rahebi. Human identification using palm print images based on deep learning methods and gray wolf optimization algorithm. *Signal, Image and Video Processing*, 18(1):961–973, 2024.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 11,2024 at 09:57:22 UTC from IEEE Xplore. Restrictions apply.

2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT) | 979-8-3503-7131-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/InCACCT61598.2024.10550972

IRIS: Intelligent Responsive Interface Solution, a promising Desktop Assistant for Differently-Abled

Aryan Karkra
Department of Information Technology
Delhi Technological University
Delhi-110042, India
aryankarkra@gmail.com

Raavi Singh
Department of Information Technology
Delhi Technological University
Delhi-110042, India
raavisingh2001@gmail.com

Dinesh K. Vishwakarma
Department of Information Technology
Delhi Technological University
Delhi-110042, India
dvishwakarma@gmail.com

Abstract*— In a quickly progressing digital world, the arrival of virtual assistants has become a revolutionary principle for human-computer interactions. In this research project, an AI Desktop Assistant inspired by Jarvis's concept is presented, with the focus on helping users through voice commands. Specifically, this initiative aims to address the requirements of people with disabilities and enhance the computing experiences. The Desktop Assistant seamlessly integrates speech recognition model, OpenAI GPT models, and other functionalities to provide extensive and user-friendly interactions. The architecture of the system is adequately designed to suitable users, comprising of speech recognition modules and models for turning text to speech. The desktop assistant model supports additional features, involving locating and managing files, playing music, moving mail, and storing the requests-responses made via OpenAI's API in the system. The system also supports surfing the web, allowing an environment that allows different users to use the technology efficiently and emphasize the safety of those users. The results reveal that the AI desktop model is running effectively in addressing the weaknesses of the existing models. Our model's capabilities include responding to commands while making the navigation and engagement of disabled persons. The system features a performance rate up to 90-95 percent, enabling interactions with users of varying dialect and accents.

Keywords*— Speech Recognition, Artificial Intelligence, Desktop Virtual Assistant, OpenAI

I. INTRODUCTION

Given that the involvement of desktops/laptops is becoming increasingly prevalent in the modern technological era, these technologies are proving to be a crucial source for work, communication, and information. Additionally, the profound and increasingly urgent need for users to have a holistic, all-in-one computing experience is made evident in the lack of inclusivity in the form of a vocal assistant that can be utilized through a PC/Laptop. As such, users and particularly those with disabilities and other special needs, who are looking to enhance their experience need a specific option to assist their journey [1]. After consideration of this vital issue, the IRIS project arose as a necessary and ambitious step into the reality of smart AI desktop assistants.

Inspired by the concept of a visionary, referred to as Jarvis or AI Assistant, the IRIS or Intelligent Responsive interface solution aims at revolutionizing desktop assistants. The IRIS has been designed to emphasize the overall accessibility, inclusivity, and computing experience for its users. This model addresses an important void, as it serves as a vital response to the growing need for solutions that accommodate people with disabilities [2]. Generally, this research project is fuelled by the current development of the intelligent virtual assistants, in particular, by providing a model that helps in managing day-to-day operations of individuals while addressing their varying capacities.

IRIS focused on the fact that it has features specific usability attributes relative to data management and search. Therefore, IRIS could handle numerous tasks by the user, from intelligent web application tasks to effective data management. Unlike some of the models that requires hard-built proprietary methods to collect data [3][4], IRIS tapped the potential of advanced speech recognition and language processing technology to provide users with authentic and seamless data search and retrieval opportunities. As such, IRIS could work as a desktop assistant that would actually perform well and function effectively for the computer.

One of the key theoretical concepts that IRIS employed is speech recognition that facilitates the user's interaction with the model through vocal commands. Furthermore, IRIS uses the Google speech-to-text engine, which offers multi-language options for data detection [5]. Finally, IRIS is connected to various APIs for quick real-time updates related to weather, news, music, journals, mails, etc [3][6]. It is crucial to note that integration of GPT-3.5-turbo-instruct from OpenAI has transformed IRIS into a conversational assistant, promoting effective interaction. Moreover, IRIS can additionally save any prompts on the desktop automatically, making it comfortable for users to reuse and refer to later. Overall, the model focuses on the features observed in other virtual assistants, including Cortana and Copilot, to overcome their drawbacks delivering a promising desktop assistant for people with disabilities.

Issues related to keeping the security of the data and privacy of the users is kept as a top priority in IRIS. The technology has been developed using the highest possible safety measures to ensure that the experience is safe and secure for users. In contrast to the other models where there can be some outside interference [7], all the functionalities and implementation in IRIS are contained in a remote level. This feature of the IRIS model allows limited control over any given flow of data . Also, it is important to state the fact that the model's focus on inclusivity naturally makes provision for its increased accessibility. The purpose of this paper is to demonstrate an in-depth examination of its architecture and functionalities to provide a detail understanding of the contribution of IRIS to the field of intelligent virtual assistants [8]. In the following sections, we will examine the structure and implication of the IRIS model, focusing on its advanced technology and aspects that are unique and value for all users, enhancing their productivity.

The research encompasses a conclusion consisting of a comparative analysis with other DVAs, reflecting on how IRIS outperforms the existent types of desktop virtual assistants such as Copilot & Cortona, in terms of its performance and unique features, specifically in relation to serving users with disabilities. The exploration of IRIS is guided by efforts in both the academic research and the pursuit of technological breakthroughs.

979-8-3503-7131-4/24/$31.00 ©2024 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 20,2024 at 06:21:41 UTC from IEEE Xplore. Restrictions apply.

II. LITERATURE REVIEW

Desktop voice assistants emerge as innovative tools that bring about rapid changes to human-computer interaction, by the means of natural-language interaction, and human task automation. The domain reveals a vivid multitude of advancements, which, nonetheless, seem to face as many challenges, as those encompassing the development from voice recognition to the speech impediments and linguistic differences or accents. Among the observed privacy concerns, security lapses and data gathering are very relevant. The necessity of understanding past work is indispensable in developing the future virtual assistants and making them a part of our everyday life. Some of the notable work done with different visions are listed below:

Inclusivity and Accessibility: Yadav et al. (2020) [1] worked on the needs of diverse users by designing their desktop assistant model for the visually impaired users, integrating screen reader assistance and customizable audio feedback. Kumar et al. (2023) [2] focused on exploring alternative input methods beyond voice commands, aiding accessibility for individuals with disabilities. These efforts ensured DVAs benefit for a large variety of users.

Context Awareness and Personalization: Biradar et al. (2023) [4] and Belkadi et al. (2020) [8] intended on personalizing the DVA experience through focusing on users and context-aware decision-making. The ability to understand individual needs and adapt responses accordingly allows new approaches for personalized assistance. However, user data plays a vital role in this aspect, potentially raising concerns about privacy and hinder the upscaling of the model (Biradar et al., 2023 [4]).

Multi-Domain Capabilities and Intelligence: Gupta et al. (2022) [3] and Mahesh (2023) [7] highlight DVAs' ability to handle complex tasks across email management, web browsing, and file manipulation (Gupta et al., 2022) [3]. So, by integration with other software, the scope of DVAs is only destined to grow, essentially transforming them into personal assistants. Furthermore, the emergence of a safe cloud-based architecture, as proposed by Mahesh (2023) [7], promotes the principles of scalability and continuous learning. Thus, by being part of a cloud-centered architecture, DVAs will access significant computing power. This will promote their ability to solve complex problems, and simultaneously, quickly and seamlessly respond to user actions and information changes.

Challenges and Drawbacks: The progress, however, is accompanied with several limitations. Their inability to grasp the true meaning of language, inaccessibility to complex tasks [1],[9], and high dependence on the user [3], make autonomous functioning of these models impossible. In addition, reliance on external services [6] and usage of cloud [7] increase the problems of data privacy and security. Currently used desktop assistants Cortona and Copilot are inefficient in using voice prompts to perform activities such as file retrieval. This requires an individual to spend a lot of time looking for files manually, which is unproductive particularly to persons with disabilities.

Future Directions: The future of desktop assistant models should concentrate on addressing these concerns whilst reaffirming their areas of strength. Further research is necessary in natural language processing to accommodate local languages and accents, decision-making algorithms, and ethical considerations. Model presented by Kunekar et al. (2023 [10]) involving an open-source base, can accelerate development and community-driven innovation. The Desktop Assistant models present a glimpse into the future of DVAs, where intelligence, personalization, and accessibility converge to create truly intelligent and user-centric assistants. Addressing the identified challenges and embracing responsible development practices will be crucial for unlocking the full potential of this technology and shaping the future of human-computer interaction. Taking into consideration the significant work done towards building better desktop assistants, our focus was to eliminate the shortcomings or the drawbacks of most of the other present models. Thus, developing a more optimized and efficient desktop assistant that ensures accessibility, inclusivity, and multi-domain operability.

III. PROPOSED METHODOLOGY

The development of the IRIS model was a meticulously planned and executed process, driven by a synthesis of innovation, research, and hands-on development. In **Figure 1,** the stages through which the model operates can be observed. Each stage is explained further, along with the parametric values used.

1. **Speech Acquisition:** The microphone captures the sound signals as the voice input. This step can be achieved by using either an external or a built-in microphone. The sound signals are then converted into speech input based on several parameters that can be configured for different users or environmental settings. **Table 1.** represents the various parameters and their values used in our model. The Energy threshold was initially kept very high at **300 decibels** in order to eliminate any noise possible and after that, it adjusts to an ambient threshold to retrieve the speech input. Other parameters were also worked with to obtain the best results.

TABLE 1. SPEECH RECOGNITION PARAMETERS

Parameter	Value
Energy_threshold	300 decibels
Dynamic_energy_adjustment_damping	0.15
Dynamic_energy_ratio	1.5
Pause_threshold	0.75 seconds
Operation_timeout	600 seconds
Phrase_threshold	0.3 seconds
Non_speaking_duration	0.5 seconds

2. **Speech Recognition Module:** The speech input was then passed into the speech recognition module where the signals are analyzed and broken into tokens based on various characteristics such as frequency, amplitude, and other patterns. The speech recognition module used was highly trained for multilingual support and allowed quick and accurate speech recognition.
3. **Command Matching/Hub:** The speech input after being converted into text is then matched with the corresponding functionality to be executed, some of them are: “search and open”, “use AI”, “news”, “play music”, “send email”, etc. In case the input doesn’t match with any of the existing tags for separate functionalities, the input is then passed onto ChatGPT using OpenAI’s API as a prompt from where a new variety of tasks can be achieved.
4. **Command Execution:** Based on the input, whichever command is triggered executes very quickly, and based on different types of commands, different kinds of output might be generated. In case of file searches, the command will print out the path of the file and even open it, web searches retrieve data and display them in an organized manner, and prompts sent to OpenAI’s API

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 20,2024 at 06:21:41 UTC from IEEE Xplore. Restrictions apply.

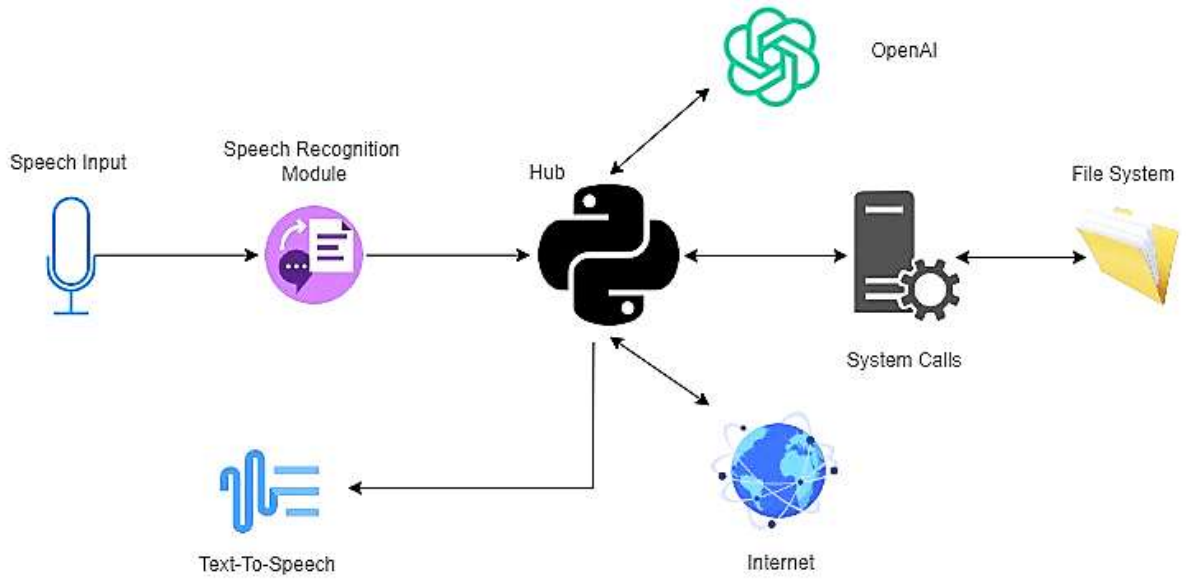


Fig.1. Model Architecture

for properly constructed responses get saved at the designated path.

5. **System Interaction:** Many of the functionalities require the model to interact with the kernel, this is done using the OS library in Python language, and many tasks such as file management including opening files, applications, etc, or retrieving environment variables for the system are fulfilled using this library.
6. **Text to Speech output:** The final output is then converted back into speech using the pyttsx3 python library, the model allows various voice settings based on the accent, volume, and speed of the output.
7. **Iterative Development and Testing:** - The development of IRIS was an iterative process, characterized by continuous testing, refinement, and user feedback. Regular iterations allowed us to address bugs, enhance functionalities, and adapt to the evolving needs of the model. User testing played a crucial role in ensuring that IRIS was not only technically robust but also user-friendly.

The model under the specified parameters was able to recognize multiple accents and languages with an appreciable accuracy of up to **90%,** the model thus delivers an impactful solution in the field of desktop virtual assistants targeted toward people with disabilities.

IV. RESULTS

The performance of our model is evidently superior to the other present models in various aspects such as accessibility, inclusivity, and multi-domain operability. The model is developed to perform many new functionalities that in turn overcome many limitations of the existing models. **Table 2.** highlights the key shortcomings present in the other models that were tackled during the development of this model.

TABLE 2. LIMITATIONS OF OTHER MODELS

PAPER/MODEL	SHORTCOMINGS
Yadav et al. (2020) [1]	Difficulty handling complex tasks, and constant wake-up calls required for the assistant.
Kumar et al. (2023) [2]	Not equally effective for all kinds of users, reducing inclusivity, unable to handle complexity.
Gupta et al. (2022) [3]	Significant user involvement and guidance are required for tasks.
Biradar et al. (2023) [4]	Heavy reliance on user data, also raises privacy concerns.
Kulhalli et al. (2017) [5]	Only provides basic commands and limited scope of usage
Sai et al. (2023) [6]	Relies on external services for tasks, thus potentially impacting performance and reliability.
Mahesh (2023) [7]	Potential security vulnerabilities, and difficulty ensuring offline functionality.
Belkadi et al. (2020) [8]	Limited features and accessibility for the users.
Co-Pilot	Difficulty accessing File systems, can't open files directly.
Cortana	Makes the overall system slower, resource demanding, and has trouble with file systems.

The limitations faced by the older models listed in **Table 2.** were worked upon and most of them were eliminated, like the model presented by Yadav et.al [1] faced challenges of the continuous need for wakeup words otherwise the instructions were not delivered, our model makes use of continuous hearing periods with an operation timeout value at **600 seconds** while keeping a high threshold value of **300 decibels** which adjusts automatically to ignore minor noises or background conversations but pick up directed instructions easily, also ensuring privacy. One of the biggest shortcomings that were common amongst almost all models was the inability to perform or respond to complex tasks, in our model we addressed this issue by seamlessly integrating it with the OpenAI API that makes use of the **gpt-3.5-turbo-instruct** model to respond to complex queries by allowing users to send custom prompts of up to **256 tokens** and also helps users to save the responses to the prompts for future use in any desired format such as JSON, .txt, etc.

The desktop assistant model IRIS is capable of recognizing various accents and languages, the model was thoroughly tested using a large sample space of people with different accents and languages which resulted in an accuracy of up to **90-95**% for the English-India language and close to **85-90%** for Hindi. File searching optimization was one of the main targets during the development of the model. Though none of the existing modern-day desktop assistant models provide an

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 20,2024 at 06:21:41 UTC from IEEE Xplore. Restrictions apply.

efficient and effective voice-controlled file retrieval functionality, Modern-day commercial desktop assistants like Cortana and Co-Pilot only allow opening root applications and not the file itself, such as PowerPoint, Adobe Reader, etc. causing a huge limitation in these models. **Table 3.** highlights the time taken by the modern-day desktop assistants to perform the same task of searching and opening a file, since only our model could open the exact file in one command, the other models opened the application required for the file. Our model outperformed the other commercially available desktop assistants as the average time taken for searching and opening the file was only **2.62 seconds** whereas the other models like Cortana took **7.16 seconds** and Co-pilot took **5.24 seconds** after an average of **10 readings** which can be seen in the graph **(Figure 2.)** as well.

TABLE 3. TIME COMPARISON FOR FILE RETRIEVAL

Runtime (seconds)	IRIS	CORTONA	COPILOT
Reading 1	2.4991	7.6022	5.3898
Reading 2	2.7731	6.8121	5.1006
Reading 3	2.5181	7.1568	5.1221
Reading 4	2.7499	6.5643	5.4167
Reading 5	2.5415	7.3389	5.3754
Reading 6	2.5732	7.4977	5.3487
Reading 7	2.7362	7.6054	5.0652
Reading 8	2.7901	7.2689	5.3391
Reading 9	2.4598	7.5771	5.4299
Reading 10	2.5260	6.9396	5.3166
Average	**2.6270**	**7.1690**	**5.2455**
Standard Deviation	**0.1103**	**0.3355**	**0.2392**

The file search algorithm was made possible with the help of the OS Python library, the search algorithm was further optimized in our model by including several changes such as case-insensitive matching and parallel searching, which also

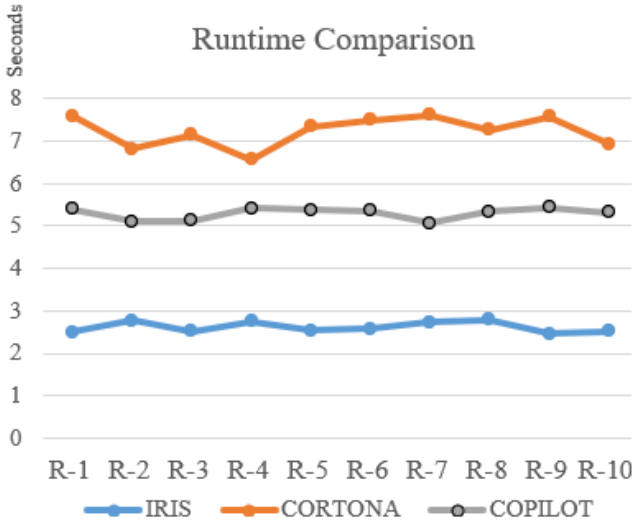


Fig.2. Runtime Comparison in File Retrieval

ensures offline functioning eliminating drawbacks faced in [7]. Many of the other applications inculcated into the model such as controlling system settings, music player, reminders, alarms, executing code scripts, file management, etc., were designed such that they function even with offline conditions. Apart from this, the model provides access to the latest news, weather reports, blogs, websites, send mails, etc. Our model increases the efficiency and quality of the work done by the users by providing AI-generated grammar corrections, text paraphrasing, language translation, math solving, etc. in desired local files such as Word, PowerPoint, Excel, etc. Thus, our model has performed better than all the existing desktop assistants in the functionality domains that have been implemented and thus it provides an optimistic signal towards further research and development.

V. CONCLUSIONS AND FUTURE WORK

In conclusion, our desktop assistant model IRIS stands as a groundbreaking AI desktop assistant, seamlessly blending advanced speech recognition technology with diverse functionalities, and engaging personality. Its commitment to inclusivity is evident in aiding individuals with disabilities and breaking down barriers to technology access. IRIS is multifunctional since it has a lot of features which make work and using it accessible. The latest and the pivotal one is considered an integration with the OpenAI's GPT-3.5-turbo-instruct model which makes possible more refined conversations. On the other hand, there is a constant update of such features as language translation and interactions with the file system within IRIS signifies the trait of adaptability and makes the technology's use diversified and productive. Apart from being an exceptional assistant, IRIS enriches lives of it users, which suggests that AI is a transforming technology. The results obtained with the model are indeed promising since IRIS calls for further research which could develop the existing applications.

The future development of the IRIS model suggests numerous opportunities for advanced interaction and functionality. Among the ideas for further development, it could be possible to consider the development of a new desktop access management system, safely integrating third-party services, users' self-created APIs for personalization, online payments and orders, and most importantly security. The development of these ideas would help in taking IRIS to a new level, allowing users to have a more effective and interesting AI desktop assistant. The development of the interaction and attraction are possible, and this is part of the next generation of intelligent virtual assistants.

VI. REFERENCES

[1]. "Desktop Voice Assistant for Visually Impaired," Regular, vol. 9, no. 2, pp. 36–39, Jul. 2020, doi: https://doi.org/10.35940/ijrte.a2753.079220.

[2]. V. Kumar, Sitanshu Barik, S. Aggarwal, D. Kumar, and V. Raj, "The use of artificial intelligence for persons with disability: a bright and promising future ahead," Disability and Rehabilitation: Assistive Technology, pp. 1–3, Dec. 2023, doi: https://doi.org/10.1080/17483107.2023.2288241.

[3]. U. Gupta, U. Jindal, A. Goel, and V. Malik, "Desktop Voice Assistant," International Journal for Research in Applied Science and Engineering Technology, vol. 10, no. 5, pp. 901–905, May 2022, doi: https://doi.org/10.22214/ijraset.2022.42390.

[4]. "PERSONAL VIRTUAL VOICE DESKTOP ASSISTANT AND INTELLIGENT DECISION MAKER," International Research Journal of Modernization in Engineering Technology and Science, 2023, doi: https://doi.org/10.56726/irjmets33892.

[5]. K. V Kulhalli, K. Sirbi, Mr. A. J. Patankar, and Research, "Personal Assistant with Voice Recognition Intelligence," International Journal of Engineering Research and Technology, vol. 10, no. 1, 2017.

[6]. P. Sai, Tripuram Pavani Kalki, P Roshini, and S. Navaneethan, "Varoka-Chatbot: An Artificial Intelligence Based Desktop Partner," Jan. 2023, doi: https://doi.org/10.1109/iceconf57129.2023.10083961.

[7]. T. R. Mahesh, "Personal A.I. Desktop Assistant," International Journal of Information Technology Research and Applications, vol. 2, no. 2, pp. 54–60, Jun. 2023, doi: https://doi.org/10.59461/ijitra.v2i2.58.

[8]. F. Belkadi, M. A. Dhuieb, J. V. Aguado, F. Laroche, A. Bernard, and F. Chinesta, "Intelligent assistant system as a context-aware decision-making support for the workers of the future," Computers & Industrial Engineering, vol. 139, p. 105732, Jan. 2020, doi: https://doi.org/10.1016/j.cie.2019.02.046.

[9]. A. Sinha, H. Muley, J. Ghosh, and Padmavati Sarode, "AI based Desktop Voice Assistant for Visually Impared Persons," Jun. 2023, doi: https://doi.org/10.1109/icces57224.2023.10192894.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 20,2024 at 06:21:41 UTC from IEEE Xplore. Restrictions apply.

[10]. Pankaj Kunekar, A. Deshmukh, Sachin Gajalwad, Aniket Bichare, Kiran Gunjal, and Shubham Hingade, "AI-based Desktop Voice Assistant," Jan. 2023, doi: https://doi.org/10.1109/icnte56631.2023.10146699.

[11]. A. Bajaj and D. K. Vishwakarma, "HOMOCHAR: A novel adversarial attack framework for exposing the vulnerability of text based neural sentiment classifiers," Eng Appl Artif Intell, vol. 126,

[12]. Nov. 2023, doi: https://doi.org/10.1016/j.engappai.2023.106815.

[13]. A. Bajaj and D. K. Vishwakarma, "A state-of-the-art review on adversarial machine learning in image classification," Multimed Tools Appl, 2023, doi: https://doi.org/10.1007/s11042-023-15883-z.

[14]. A. Bajaj and D. Kumar Vishwakarma, "Evading text based emotion detection mechanism via adversarial attacks," Neurocomputing, vol. 558, p. 126787, Nov. 2023, doi: https://doi.org/10.1016/j.neucom.2023.126787.

[15]. A. Bajaj and D. K. Vishwakarma, "Exposing the Vulnerabilities of Deep Learning Models in News Classification," in 2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT), IEEE, Feb. 2023, pp. 1–5. doi: https://doi.org/10.1109/ICITIIT57246.2023.10068577.

[16]. A. Bajaj and D. K. Vishwakarma, "Bypassing Deep Learning based Sentiment Analysis from Business Reviews," in 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN), IEEE, May 2023, pp. 1–6. doi: https://doi.org/10.1109/ViTECoN58111.2023.10157098.

[17]. A. Verma, Virender Ranga, and Dinesh Kumar Vishwakarma, "A novel approach for forecasting PM2.5 pollution in Delhi using CATALYST," Environmental Monitoring and Assessment, vol. 195, no. 12, Nov. 2023, doi: https://doi.org/10.1007/s10661-023-12020-z.

[18]. A. Verma, Virendar Ranga, and Dinesh Kumar Vishwakarma, "Forecasting of Satellite Based Carbon-Monoxide Time-Series Data Using a Deep Learning Approach," Feb. 2023, doi: https://doi.org/10.1109/icitiit57246.2023.10068609.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 20,2024 at 06:21:41 UTC from IEEE Xplore. Restrictions apply.

Educational Administration: Theory and Practice
2024, 30(5), 13145-13156
ISSN: 2148-2403
https://kuey.net/

Research Article

Key Factors in Automobile Design: Expert Views on Enhancing User Experience during New Product Development

Shahid Ahmad[1*], Ranganath M Singari[2], S.L. Bhandarkar[3]

[1*,2]Department of Design, Delhi Technological University, Delhi, India*shahidahd10@gmail.com
[3]Guru Gobind Singh Indraprastha University, Delhi

Citation Shahid Ahmad, et al (2024), Key Factors in Automobile Design: Expert Views on Enhancing User Experience during New Product Development, Educational Administration: Theory and Practice, 30(5), 13145-13156
Doi: 10.53555/kuey.v30i5.5676

ARTICLE INFO

ABSTRACT

This study provides a comprehensive review of the key elements driving modern car design. The study aims to offer crucial perspectives for designers of automobiles by pinpointing important factors in several aspects of user experience design. The paper highlights the essential variables that effect the design of vehicle exteriors and interiors, human-machine interfaces, safety features, connectivity, and other crucial elements through the use of extensive surveys and in-depth talks with over 20 industry experts. This in-depth exploration of these foundational issues is intended to function as a comprehensive manual for designers tasked with the difficult undertaking of producing user-centered car designs.

Keywords: Automobile Design, User Experience, Product Development, Key Factors

Introduction

In the pursuit of creating innovative and competitive vehicles, designers and engineers grapple with a myriad of considerations, navigating the delicate balance between form and function. The field of automotive design is influenced by a myriad of factors that have evolved over time. From the early days of the horseless carriage to the present, influences such as new technology, lifestyle, culture, and consumer preferences have played a pivotal role in shaping the aesthetics and functionality of automobiles (Muscad, 2023). In the contemporary context, the focus on user experience design has become increasingly important, with factors such as safety, regulatory requirements, technological advancements, and ergonomic considerations significantly impacting the design process. As automotive design continues to evolve, designers are faced with the challenge of integrating new materials, safety features, and infotainment systems while ensuring that the design remains adaptable and complies with changing regulations and customer needs. Therefore, this research paper seeks to provide a comprehensive analysis of the essential factors that every designer must consider when designing a new automobile, considering the dynamic nature of the automotive industry, the evolving demands of consumers, and regulatory standards (Design Considerations - Automotive OEM & Tier | 3M US, n.d.).

This study endeavours to conduct a comprehensive analysis of the factors that significantly influence automobile design, with a particular focus on the pivotal role of user experience design in shaping the contours of new product development. In a world where vehicles are no longer just modes of transportation but extensions of users' lifestyles, understanding the intricate nuances of design becomes paramount.

The factors influencing automotive design are rooted in the dynamic and multifaceted nature of the automotive industry. External factors such as political, economic, social, technological, legal, and environmental influences play a pivotal role in shaping the design of automobiles. Additionally, the history of automotive design has been marked by significant engineering advances, technological shifts, and changes in consumer preferences, making it a continuously evolving field (Jaafarnia & Bass, 2011). Moreover, the influence of different markets around the world on automotive design highlights the need to understand and adapt to varying consumer tastes, climates, and historical trends.

Copyright © 2024 by Author/s and Licensed by Kuey. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The significance of this study lies in its potential to provide valuable insights for designers, manufacturers, and investors in the automotive industry (Ilić et al., 2024). By comprehensively analysing the essential factors that every designer must consider in automobile design, the study aims to contribute to the development of vehicles that not only meet regulatory standards and consumer preferences but also align with the rapidly changing technological landscape and environmental considerations. Understanding the historical and global influences on automotive design can help stakeholders anticipate future trends and make informed decisions in the design and production of automobiles (Vicky, 2023). Therefore, this study seeks to shed light on the interconnectedness of various factors influencing automotive design, ultimately aiming to enhance the quality, safety, and sustainability of vehicles in the ever-evolving automotive landscape.

The primary objectives of this research are:
To provide a comprehensive analysis of the factors that every designer must consider when designing a new automobile (Papadaki, n.d.).
To highlight the importance of a user-cantered approach in automobile design and the need for designers to consider various aspects such as safety, regulatory requirements, technological advancements, and ergonomics.

Extensive surveys and discussions with 20+ industry experts, focusing on their perspectives on user experience design factors in automobile design (Papadaki, n.d.) has been done. The results of the surveys and interviews to identify common themes and key factors influencing automobile design (Mohammed et al., 2017) have been analysed.
Reviewing literature on automobile design, including consumer user experience design factors has been the main focus of this study.
The research aims to provide a thorough understanding of the factors that designers must consider while designing a new automobile, ultimately contributing to the development of vehicles that meet consumer preferences and the rapidly changing technological landscape.

Literature Review

Automobile design has undergone a profound evolution since the inception of the automotive industry. This section of the research delves into the existing body of literature, tracing the historical trajectory of automobile design and exploring contemporary perspectives on the key factors influencing this dynamic field (Mohammed et al., 2017).

The history of automotive design is a fascinating journey that has been shaped by various influences and technological advancements. In the early days of the automotive industry, the line between art and design was not clearly defined, and vehicles were expected to balance aesthetic, market, performance, functional, and legislative considerations (Erdal, 2023). Since the inception of automobile manufacturing, influences such as new technology, lifestyle, culture, and consumer preferences have played a significant role in shaping the aesthetics and functionality of automobiles.

One of the significant developments in car design came in the 1920s and 1930s with the introduction of streamlining, inspired by the aerodynamic shapes of airplanes. Subsequently, the 1960s and 1970s saw cars featuring more futuristic designs and advanced technology, while in recent years, there has been a renewed focus on sustainability and eco-friendliness, leading to the integration of hybrid or electric powertrains and designs prioritizing aerodynamics and fuel efficiency (Kunze, 2021).

The evolution of the automotive design process has also been marked by significant changes, from an entirely manual craft to an almost entirely digital craft, with the introduction of new tools and techniques to enhance the design process. Throughout history, automotive design has mirrored the evolution of society, with changing lifestyle choices impacting design paradigms and consumer preferences influencing the direction of car design.

Below Figure 1, the steps involved in the automobile car design industry are illustrated. This figure outlines the comprehensive process, from initial concept development and market research to detailed design phases, prototyping, testing, and final production. Each step is crucial for ensuring that the final product not only meets aesthetic and functional requirements but also adheres to safety standards and user experience expectations.

Figure 1. The steps involved in the automobile car design industry are illustrated

History of automotive design is a dynamic and multifaceted narrative, influenced by a wide array of factors such as technology, culture, consumer preferences, and regulatory standards. This evolution has led to the creation of vehicles that are not only functional but also visually appealing, safe, and environmentally sustainable.

The roots of automobile design can be traced back to the early 20th century, when vehicles were primarily utilitarian, devoid of the aesthetic considerations that define contemporary designs. Over time, the automotive industry witnessed a paradigm shift from the functional designs of the Model T era to the sleek and aerodynamic forms of the mid-20th century. The incorporation of design elements, such as streamlining and tail fins, reflected not only aesthetic considerations but also the influence of societal trends and technological advancements (Kunze, 2021)..

In recent decades, the integration of advanced materials, manufacturing processes, and digital technologies has further revolutionized automobile design. The advent of computer-aided design (CAD) has allowed designers to visualize and iterate complex designs more efficiently, contributing to the sophistication of modern vehicles.

Advancements in technology, including artificial intelligence, connectivity, and electrification, have reshaped the automotive design landscape. The literature highlights the role of technology as both an enabler and a disruptor. The integration of smart technologies, such as infotainment systems, autonomous driving features, and electric powertrains (Sajjad, 2024), necessitates a re-evaluation of design principles to accommodate new functionalities while maintaining a cohesive user experience . The rise of electric and autonomous vehicles introduces novel challenges related to the reimagining of vehicle interiors, the elimination of traditional design elements like the internal combustion engine, and the exploration of innovative materials for sustainable manufacturing.

Synthesizing the insights from the literature, it becomes evident that the evolution of automobile design is a dynamic interplay between historical legacies, technological advancements, and the ever-changing landscape of consumer expectations. This foundation sets the stage for a deeper exploration of the specific factors influencing contemporary automobile design, with a keen focus on the role of user experience design in new product development (Rarecars & Rarecars, 2023). The subsequent sections will delve into the theoretical framework that guides this study, shedding light on the principles and methodologies that underpin the intricate process of designing automobiles in the 21st century.

Methodology

The research methodology employed in this study is a mixed-methods approach, combining quantitative and qualitative methods to provide a comprehensive understanding of the factors influencing automobile design,

with a specific focus on user experience (UX) design. The methodology encompasses surveys and subjective question-answer discussions with industry experts, aiming to triangulate insights and achieve a nuanced perspective on contemporary design practices.

Survey Design: A purposive sampling method was employed to select participants for the survey. Professionals with expertise in automobile design, engineering, and user experience were targeted. A diverse group of respondents, representing various roles within the automotive industry, including designers, engineers, and product managers, were included to capture a holistic view of the design process. Twenty industry experts were selected based on their extensive experience and expertise in automobile design, Colour material Finish, Modelling (Digital and Physical Modelling). The experts represented a diverse range of perspectives, including professionals from leading automotive companies, design in start-up, and academic institutions.

Subjective question-answer discussions were conducted through semi-structured interviews. The interviews were designed to elicit detailed insights into the challenges, priorities, and future trends in automobile design. Topics included the role of user experience in the design process, integration of emerging technologies, and the balance between aesthetics and functionality. These discussions were audio-recorded and transcribed for thorough analysis.

Questionnaire Structure: The survey questionnaire was designed to gather both quantitative and qualitative data. Multiple-choice questions were utilized to collect quantitative data on participants' perceptions of the importance of different design factors (Wiss et al., 2018), while open-ended questions allowed respondents to provide in-depth insights into their experiences and perspectives.

The survey covered key aspects such as aesthetic considerations, technological integration, safety features, environmental sustainability, and user experience design. Likert scales were employed to assess the degree of importance attributed to each factor, providing quantifiable data for analysis.

List of questions for survey listed below:

1. Designer Name
2. Experience
3. Company Names
4. Why do you design?
5. What do you think when you start creating new designs for cars, Bikes, and other objects?
6. What are the fundamental elements/factors that inspire you to create a design? (one word/Sentences)
7. How important is the user in automobile design process? (Scale 1-10)
8. When you start sketching/Developing new automobile/Product. You think about user?

(a) Comfort
(b) Aesthetics
(c) Feasibility
(d) Ease of use
(e) _____

4. Factors Influencing Automobile Design:

Several factors influence automotive design, encompassing a wide range of considerations that shape the appearance, functionality, and regulatory compliance of motor vehicles. These factors include:
Consumer Preferences: Consumer preferences play a pivotal role in automotive design. Designers must consider factors such as styling, functionality, and features that appeal to consumers and differentiate their vehicles from competitors (Muscad, 2023).

Safety and Regulatory Requirements:

Safety considerations are paramount in automobile design, adhering to both industry regulations and consumer expectations. Survey participants emphasized the importance of safety features such as advanced driver assistance systems (ADAS), crashworthiness, and collision avoidance technologies. Ensuring that vehicles meet safety standards and regulatory requirements—including crash tests, emissions standards, and other mandated regulations—is a crucial aspect of automotive design (Muscad, 2023).

The survey indicated a high level of importance attributed to safety features, reflecting the industry's commitment to enhancing vehicle safety standards. Expert discussions underscored the ongoing challenges and opportunities in integrating evolving safety technologies while maintaining a focus on overall design coherence.

Technological Integration:

Advancements in technology have significantly influenced automobile design, as highlighted in both the survey and expert discussions. The integration of cutting-edge technologies, including infotainment systems, connectivity features, and driver assistance technologies, is considered paramount in contemporary vehicle design. It includes innovations in materials, electronics, and software, significantly impact automotive design. Designers must integrate these advancements to enhance the functionality and appeal of vehicles.

Survey results revealed a consensus among respondents regarding the high importance placed on technological integration in the design process. Industry experts emphasized the need to strike a balance between incorporating innovative technologies and ensuring user-friendly interfaces that do not overwhelm users with complexity.

Ergonomics and Human Factors:

The ergonomic design of automobile interiors, focusing on user comfort, safety, and ease of use, emerged as a critical factor in both survey responses and expert discussions. Human factors engineering principles, encompassing aspects such as seating position, visibility, and control accessibility, were highlighted as crucial for enhancing the overall user experience. It is a critical factor in automotive design, encompassing seating arrangements, cargo capacity, and other functional aspects that impact the user experience. Designers must prioritize the comfort and usability of vehicles.

Survey participants emphasized the importance of ergonomic considerations in designing controls, interfaces, and seating arrangements. The integration of human-centric design principles aims to create vehicles that minimize driver fatigue, enhance safety, and provide a comfortable and intuitive driving experience (Jaafarnia, 2019).

Aesthetic Considerations:

Aesthetics play a pivotal role in influencing consumer preferences and perceptions of a vehicle. Survey results indicate that a significant majority of respondents rank aesthetic considerations, such as exterior design, interior styling, and overall visual appeal, as highly important in the design process. Industry experts, during subjective discussions, emphasized the significance of creating visually striking and distinctive designs to captivate consumers in a competitive market (Estrela, 2023).

Designers often strive to strike a balance between innovation and maintaining brand identity, ensuring that the aesthetic appeal aligns with the target market's tastes. Aesthetic considerations extend beyond visual appeal, encompassing elements like colour choices, material selection, and design coherence throughout the vehicle. The design of vehicles must balance aesthetics and functionality, considering factors such as aerodynamics, safety, and the overall appearance of the vehicle (Liamadis & Tsinikas, 2013).

Environmental Sustainability:

Environmental sustainability emerged as a notable factor influencing contemporary automobile design. Survey participants expressed a growing awareness of environmental concerns, with a significant proportion indicating that factors such as fuel efficiency, emissions reduction, and the use of eco-friendly materials are important considerations in the design process. The modern automobile design must incorporate requirements for pollution-control components, reflecting the industry's commitment to environmental sustainability (Wellbrock et al., 2020).

Experts discussed the increasing emphasis on sustainability in the automotive industry, with a focus on developing eco-friendly vehicles, exploring alternative propulsion systems, and adopting sustainable manufacturing practices. The integration of environmental sustainability considerations reflects a broader societal shift towards greener and more responsible design practices.

Automotive design is shaped by a multitude of factors, including consumer preferences, regulatory requirements, technological advancements, and historical influences. Designers must skilfully navigate these considerations to develop vehicles that are visually appealing, safe, functional, and compliant with industry standards (Duragappa & Sheshadri, 2020). The design process is a complex interplay of these elements, collectively determining the form, functionality, and user experience of a vehicle. This section delves into the key factors influencing automobile design, integrating insights from quantitative survey data and qualitative discussions with industry experts.

Study of various models

Examining case studies provides a practical lens through which to understand how user experience (UX) design principles have been applied in the development of automobiles. The following case studies illustrate successful implementations of UX design in the automotive industry, showcasing the integration of user-centric considerations in the design process.

Tesla Model 3: Elevating User Experience through Simplicity and Innovation:

The Tesla Model 3, an all-electric sedan, exemplifies a user-centric approach to automobile design. Tesla's commitment to innovation, coupled with a focus on simplicity and user-friendly interfaces, sets the Model 3 apart in the automotive market. The Tesla Model 3 is widely regarded as one of the best cars in terms of user experience due to its remarkable blend of simplicity and innovation (Bluestone, 2021). The car's minimalist design approach, which eliminates traditional physical controls and relies heavily on a single large touchscreen, creates a seamless and intuitive interface that empowers drivers. The Model 3's advanced technology, including AI-driven features and over-the-air updates, ensures that the user experience is constantly evolving and improving. Additionally, the car's focus on simplicity and ease of use makes it accessible to a wide range of users, from those new to electric vehicles to seasoned tech-savvy drivers (Miccio, 2020). Overall, the Tesla Model 3's innovative design and user-centric approach have set a new standard for the automotive industry, making it a benchmark for other manufacturers to follow. Below figure2 the image wallpaper of Tesla Model 3.

Figure 2. Tesla Model 3

Key UX Design Features:

Minimalist Interior: The Model 3's interior is characterized by a minimalist design, featuring a large central touchscreen as the primary interface. This design choice eliminates physical buttons and switches, creating a clean and uncluttered dashboard.

Intuitive Controls: The touchscreen interface consolidates controls for navigation, entertainment, climate, and vehicle settings, providing users with an intuitive and centralized control hub. Voice commands further enhance accessibility.

Over-the-Air Updates: Tesla's unique capability to deliver software updates over-the-air enables continuous improvement of features, addressing user feedback and introducing new functionalities without requiring a physical visit to a service centre.

Impact and Results: The Model 3's user-friendly interface and continuous software updates contribute to a positive and evolving user experience.

The emphasis on simplicity aligns with the preferences of modern consumers, and the streamlined design has become a hallmark of Tesla's UX philosophy. Tesla's UX philosophy is rooted in a forward-thinking approach that emphasizes simplicity, innovation, and user-centric design. From its early years, Tesla has prioritized creating an intuitive and seamless user experience, integrating cutting-edge technology with a minimalist aesthetic. Central to this philosophy is the belief that the car should be an extension of the driver's digital life, offering a seamless interface that is both functional and elegant. Tesla's continuous software updates and over-the-air enhancements reflect a commitment to evolving and improving the user experience, ensuring that each vehicle becomes better over time. The large touchscreen interface, advanced driver assistance systems, and focus on automation and connectivity illustrate Tesla's dedication to redefining how users interact with their vehicles, setting a new standard for automotive UX in the industry.

Volvo XC40 Recharge: Prioritizing Safety and Sustainability in UX Design:

Volvo, renowned for its commitment to safety, has extended its principles to the XC40 Recharge, an electric SUV. The XC40 Recharge integrates user experience design with a strong emphasis on safety and sustainability. The Volvo XC40 Recharge stands out as one of the best cars in terms of prioritizing safety and sustainability in UX design. Volvo's commitment to safety is deeply ingrained in the XC40 Recharge, evident through its advanced safety features and robust construction that prioritize the well-being of both passengers and pedestrians (D'souza & Overdrive.In, 2022). Additionally, the car's sustainability focus is showcased through its electric powertrain, emphasizing Volvo's dedication to reducing environmental impact. In terms of user experience design, the XC40 Recharge seamlessly integrates safety and sustainability features into its user interface, ensuring that drivers not only enjoy a luxurious and comfortable ride but also feel secure and environmentally conscious while driving. This holistic approach to safety, sustainability, and user experience design positions the Volvo XC40 Recharge as a standout choice for those seeking a vehicle that excels in all these crucial aspects (Volvo SUVs: The Perfect Blend of Luxury, Safety, and Sustainability, n.d.). Below figure 3 and 4 show the wallpaper of Volvo XC40 Recharge of interior and exterior.

Figure 3. Volvo XC40 Recharge interior

Figure 4. Volvo XC40 Recharge exterior

Key UX Design Features:
Safety-Centric Design: Volvo's City Safety features, including collision avoidance systems and pedestrian detection, prioritize user safety. The interior design incorporates features like run-off road protection and whiplash protection.

Seamless Charging Experience: The XC40 Recharge focuses on simplifying the electric vehicle (EV) charging experience. The infotainment system provides information on charging station locations, charging times, and remaining battery range.
Sustainable Materials: The interior of the XC40 Recharge uses sustainable materials, contributing to Volvo's commitment to environmentally conscious design.
Impact and Results: The integration of safety features aligns with Volvo's brand identity, appealing to consumers who prioritize safety in their vehicle choices.

The UX design considerations for electric vehicle charging contribute to a more accessible and convenient experience for users adopting electric mobility.

BMW iX: Merging Luxury with Sustainable Innovation in UX Design:
The BMW iX represents BMW's foray into sustainable luxury with an emphasis on electric mobility. This case study highlights the integration of user experience design principles in creating a futuristic and environmentally conscious vehicle. The BMW iX stands out as one of the best cars in terms of merging luxury with sustainable innovation in UX design. The iX's interior showcases a stunning new design language that replaces traditional plastics with high-quality materials like crystal, wood, and fabrics (10 Best Interiors & UX Winner: BMW iX, 2024). This commitment to sustainability extends to the use of recycled materials throughout the cabin, including Econyl from recycled fishing nets and olive leaf-tanned leather (BMW iX Named a 2022 Wards 10 Best Interiors and UX Winner | Uncategorized, 2024).At the same time, the iX's interior exudes a sense of luxury and craftsmanship, with elegant aesthetics, warm colors, and innovative design elements. The minimalist dashboard is dominated by the BMW Curved Display, a sweeping glass panel that houses the 12.3-inch digital instrument cluster and 14.9-inch infotainment screen (Salem, 2022). This seamless integration of technology and premium materials creates a calming, lounge-like atmosphere that prioritizes the user experience. The iX's advanced features, such as the intelligent personal assistant and smart fabrics, are discreetly integrated into the design, allowing the driver to focus on the joy of driving (Banks, 2021). By striking a balance between sustainability, luxury, and cutting-edge technology, the BMW iX sets a new standard for the future of automotive design and user experience. Below figure 5 and 6 show the wallpaper of BMW iX of interior and exterior.

Figure 5. BMW iX interior

Figure 6. BMW iX exterior

Key UX Design Features:

Cohesive Digital Cockpit: The iX features a digital cockpit with a large curved display, incorporating controls for navigation, entertainment, and vehicle settings. The interface is designed for intuitive interaction, minimizing distractions for the driver.

Augmented Reality Head-Up Display: BMW's Head-Up Display integrates augmented reality elements, providing navigation instructions, safety alerts, and other information directly in the driver's line of sight.

Eco-Friendly Materials: The iX incorporates sustainable materials in its interior, reflecting BMW's commitment to reducing the environmental impact of its vehicles.

Impact and Results: The cohesive digital cockpit enhances the overall user experience, offering a blend of advanced technology and user-friendly interfaces.

BMW's focus on sustainability in materials aligns with the growing consumer demand for eco-friendly choices in luxury vehicles.

The case studies of leading automotive manufacturers exemplify how they leverage UX design principles to prioritize user needs, safety, sustainability, and innovation in their vehicles. By integrating these crucial considerations into their design processes, these companies have not only elevated the user experience but also established new benchmarks for the industry's evolution. Through a focus on user-centric design, safety features, sustainable materials, and cutting-edge technology, these manufacturers have redefined what it means to create a truly exceptional driving experience. By placing the user at the centre of their design philosophy and incorporating the latest advancements in safety and sustainability, these companies have not only met but exceeded consumer expectations, setting a high standard for the automotive industry as it progresses into the future.

Result

Here is a table 1 summarizing the key user experience design factors that impact automobile design, based on the survey and case studies results:

Table1: UX Design Factors thst impaxr Automobile Design

UX Design Factor	Impact on Automobile Design
Safety	- Designers must ensure vehicles meet safety standards and regulatory requirements
	- Safety features and advanced driver-assist technologies are integrated to enhance occupant protection
Functionality	- Users prefer functions related to functionality, especially safety and assistance in driving
	- Functionality is prioritized over entertainment to achieve a high level of user experience
Ergonomics	- Designers must prioritize user comfort, convenience, and safety by focusing on seating arrangements, controls placement, and overall user experience
Aesthetics	- Aesthetic considerations influence the visual appeal and brand identity of vehicles
	- Designers must balance aesthetics with functionality to create visually appealing and distinctive designs
Sustainability	- Eco-friendly materials and lightweight construction techniques are used to reduce environmental impact
	- Seamless charging experience and energy-efficient technologies are integrated for electric vehicles
Connectivity	- Advanced infotainment systems, digital cockpits, and augmented reality features provide a seamless and immersive driving experience
Ease of Use	- Intuitive controls and minimalist interior design enhance the ease of use and user experience
Accessibility	- Designers must ensure that vehicles are accessible to users with diverse needs and abilities
Responsiveness	- The vehicle's systems and interfaces should respond promptly to user inputs and commands
Navigation	- Clear and intuitive navigation systems help users easily find their way and access desired features

The key user experience design factors that directly impact automobile design include safety, functionality, ergonomics, aesthetics, sustainability, connectivity, ease of use, accessibility, responsiveness, and navigation. Designers must ensure vehicles meet safety standards and regulatory requirements, prioritize user comfort and convenience, and balance aesthetics with functionality to create visually appealing and distinctive designs. Additionally, eco-friendly materials and lightweight construction techniques are used to reduce environmental impact, while advanced infotainment systems and digital cockpits provide a seamless and immersive driving experience. Intuitive controls and minimalist interior design enhance ease of use, while accessibility ensures that vehicles cater to users with diverse needs and abilities. Responsiveness and navigation are also crucial, with the vehicle's systems and interfaces responding promptly to user inputs and providing clear and intuitive navigation systems.

Conclusion

The evolution of automobile design is intricately linked to the dynamic interplay of technological advancements, changing consumer expectations, and the integration of user experience (UX) design principles. This research has delved into the multifaceted landscape of automobile design, exploring factors that influence the process, the role of UX design in new product development, insights from industry experts, and recommendations for future design processes.

The comprehensive analysis of factors influencing automobile design revealed the significance of safety, functionality, ergonomics, aesthetics, sustainability, connectivity, ease of use, accessibility, responsiveness, and navigation. User experience design emerged as a pivotal factor, shaping the contours of contemporary vehicle designs to meet the expectations and preferences of an increasingly sophisticated consumer base.

Insights from industry experts provided valuable perspectives on design priorities, challenges, and future directions. The emphasis on user-centricity, the challenges posed by technological advancements, and the potential for innovation through interdisciplinary collaboration were recurring themes. The case studies illustrated successful applications of UX design principles in real-world scenarios, showcasing how leading manufacturers prioritize simplicity, safety, sustainability, and innovation.

As the automotive industry continues to undergo transformative shifts, the implications for future automobile design are profound. The integration of UX design principles will be paramount in creating vehicles that not only meet functional requirements but also resonate with users on emotional and experiential levels. The recommendations put forth emphasize the need for a holistic and collaborative approach, leveraging emerging technologies and continuous user feedback to drive innovation.

The future of automobile design is poised for a paradigm shift, marked by a convergence of technology, sustainability, and user-centric innovation. Vehicles are evolving beyond mere modes of transportation, becoming personalized, connected, and sustainable extensions of users' lifestyles. Designers are tasked with

navigating a landscape where the traditional boundaries between automotive and tech industries blur, creating opportunities for ground breaking advancements in user experience and mobility solutions.

The integration of artificial intelligence, augmented reality, and sustainable materials will play pivotal roles in shaping the future of automobile design. As the industry embraces these advancements, a delicate balance between functionality and aesthetics, safety and innovation, and environmental responsibility will be crucial in creating vehicles that stand at the forefront of technological and experiential excellence. The journey through the intricacies of automobile design reveals a vibrant landscape where innovation, user experience, and sustainability converge.

References:

1. Muscad, O. (2023, March 29). What is Automotive Design? An Overview. DATAMYTE. https://datamyte.com/blog/automotive-design/
2. Design Considerations - Automotive OEM & Tier | 3M US. (n.d.). MMM-ext. https://www.3m.com/3M/en_US/oem-tier-us/automotive-interiors/future-design-considerations/
3. Jaafarnia, M., & Bass, A. (2011). Tracing the Evolution of Automobile design: Factors influencing the development of aesthetics in automobiles from 1885 to the present. *International Conference on Innovative Methods in Product Design*.
4. Ilić, T., Peštović, K., Saković, D., & Rađo, D. (2024). Evaluation of Earnings Management in Relation to the Quality of Financial Reporting: A Comparative Analysis of the Agriculture and Manufacturing Sectors. *Contemporary Agriculture*, *73*(1–2), 119–128. https://doi.org/10.2478/contagri-2024-0015
5. Vicky. (2023, April 14). *Automotive Design Process: From Concept Ideas to Production - APW*. AutoProtoWay. https://autoprotoway.com/automotive-design-process/
6. Papadaki, P. (n.d.). User experience in automotive industry: user perspective on functionality and entertainment. In *https://lnu.diva-portal.org*. Department of Informatics.
7. Mohammed, A., Hussain, I. M., & Zain, Z. B. (2017). Analysis For Determining Factors That Make Good Design Process In Automotive Manufacturing Organization. *IOSR Journal of Mechanical and Civil Engineering*, *14*(3), 97–104. https://doi.org/10.9790/1684-14030597104
8. Erdal, C. (2023, April 6). *A Brief History of Car Design*. Pedal Commander. https://pedalcommander.com/blogs/garage/a-brief-history-in-car-design.
9. Kunze, J. (2021, December 16). *The history of automotive design*. flyingshapes. https://www.flyingshapes.com/history-of-automotive-design
10. Sajjad, S. (2024, May 12). *How technology has changed the automotive industry*. https://www.linkedin.com/pulse/how-technology-has-changed-automotive-industry-saba-sajjad/
11. Rarecars, & Rarecars. (2023, October 15). *Evolution of Automobile Design*. Rarecars.com. https://rarecars.com/evolution-of-automobile-design/
12. Rarecars, & Rarecars. (2023, October 15). *Evolution of Automobile Design*. Rarecars.com. https://rarecars.com/evolution-of-automobile-design/
13. Jaafarnia, M. (2019). Evolution of Automobile Design. A Brief Introductio. *Yadegar Omr Publication*. https://www.researchgate.net/publication/336148905_Evolution_of_Automobile_Design_A_Brief_Introduction
14. Estrela, H. (2023, August 24). *The Evolution of Car Design: Balancing Aesthetics and Functionality*. https://www.linkedin.com/pulse/evolution-car-design-balancing-aesthetics-hugo-estrela/
15. Liamadis, G., & Tsinikas, N. (2013). *Automotive Design Culture: Aesthetic Trends Originated in Technology*. http://ikee.lib.auth.gr/record/216133
16. Wellbrock, W., Ludin, D., Röhrle, L., & Gerstlberger, W. (2020). Sustainability in the automotive industry, importance of and impact on automobile interior – insights from an empirical survey. *International Journal of Corporate Social Responsibility*, *5*(1). https://doi.org/10.1186/s40991-020-00057-z
17. Duragappa, N., & Sheshadri, K. N. (2020). How KOHA OPAC can be Designed and used as Library Website: A Practical Approach. *SRELS Journal of Information Management/SRELS Journal of Information Management*, 63–72. https://doi.org/10.17821/srels/2020/v57i2/149470
18. Bluestone, D. (2021, December 16). UX takeaways from Tesla's Model 3 interface - UX Collective. *Medium*. https://uxdesign.cc/ux-takeaways-from-teslas-model-3-interface-61dcc23478ab
19. Miccio, C. (2020, August 3). *UX: Simplicity is Best!* https://www.linkedin.com/pulse/ux-simplicity-best-christian-miccio/
20. D'souza, B., & Overdrive.In. (2022, April 6). 2022 Volvo XC40 Recharge first drive review. *Overdrive*. https://www.overdrive.in/reviews/2022-volvo-xc40-recharge-first-drive-review/
21. D'souza, B., & Overdrive.In. (2022, April 6). 2022 Volvo XC40 Recharge first drive review. *Overdrive*. https://www.overdrive.in/reviews/2022-volvo-xc40-recharge-first-drive-review/
22. *BMW iX Named a 2022 Wards 10 Best Interiors and UX Winner | Uncategorized*. (n.d.). https://www.passportbmw.com/blogs/846/uncategorized/bmw-ix-named-a-2022-wards-10-best-interiors-and-ux-winner/

23. Salem, B. O. (2022, June 30). *BMW iX named a Best Interior and Best UX winner by WardsAuto | BMW of Salem*. https://www.bmwofsalem.com/blog/2022/june/30/bmw-ix-named-a-best-interior-and-best-ux-winner-by-wardsauto.htm
24. Banks, N. (2021, July 19). This Is What BMW's Design Boss Adrian Van Hooydonk Feels About The Electric iX. *Forbes*. https://www.forbes.com/sites/nargessbanks/2020/11/11/this-is-the-design-story-behind-bmws-all-new-flagship-electric-ix/?sh=6bfdfb0c1469

Physica Scripta

INCLUSIVE PUBLISHING
TRUSTED SCIENCE

PAPER

Laboratory simulation to study the nonlinear evolution of right-hand circularly polarised wave in laser-produced plasmas

To cite this article: Aditya *et al* 2024 *Phys. Scr.* **99** 075605

View the article online for updates and enhancements.

You may also like

- Investigation of the Langdon effect on the nonlinear evolution of SRS from the early-stage inflation to the late-stage development of secondary instabilities
Jie Qiu, Liang Hao, Lihua Cao et al.
- Resonance Raman spectroscopy of twisted interfaces in turbostratic multilayer graphene
A Mohapatra, S Poudyal, M S Ramachandra Rao et al.
- Quantum liquids
H R Glyde

This content was downloaded from IP address 14.139.251.98 on 13/06/2024 at 08:07

Physica Scripta

PAPER

Laboratory simulation to study the nonlinear evolution of right-hand circularly polarised wave in laser-produced plasmas

Aditya[1], **Suresh C Sharma**[1], **Indraj Singh**[2] and **R P Sharma**[2]

[1] Delhi Technological University, Shahbad Daulatpur, Delhi-110042, India
[2] Indian Institute of Technology Delhi, Hauz Khas, New Delhi-110016, India

E-mail: suresh321sharma@gmail.com

RECEIVED
31 January 2024

REVISED
18 May 2024

ACCEPTED FOR PUBLICATION
28 May 2024

PUBLISHED
10 June 2024

Keywords: laboratory simulations, turbulence, R wave, nonlinear evolution

Abstract

To simulate the astrophysical scenarios at the scale of laboratory astrophysics, the interaction of high-power lasers with plasmas is increasingly being used. So, in this perspective, a nonlinear wave-wave interaction between right-hand circularly polarised wave (R-mode wave) and electron plasma wave (EPW) is proposed to understand the localization of the pump R-mode wave and associated magnetic turbulence at electron scale in magnetized plasma relevant to laboratory astrophysics. The ponderomotive force and relativistic variation of electron mass are accountable for this nonlinear wave-wave interaction model. For this nonlinear dynamics, coupled model equations for EPW and R-mode wave are developed and solved by laboratory simulations utilizing finite difference and pseudo-spectral methods. The observed simulation findings, like the nonlinear evolution of the pump wave, typical scale size of the formed localized structures, and electron scale turbulent spectra associated with the coherent structures resemble various laboratory and space plasma scenarios. A semi-analytical model in the paraxial limits is also developed to understand the physics behind the localization process of the pump wave inside the plasma medium.

1. Introduction

The invention of high-intensity lasers opened new areas of research at the scale of laboratory astrophysics concerning a wide range of magnetized plasma systems from space to astrophysical phenomena [1–3]. In various theoretical and experimental investigations of intense laser-plasma interactions in magnetized plasma, numerous features of space and astrophysical situations, such as turbulence generation, magnetic field enhancement, filamentary structure formation, particle acceleration, laser wakefield acceleration [4, 5], and so on, have been observed at the scale of laboratory astrophysics [6–14]. One of the most vital features of magnetized plasma systems is turbulence generation at electron and ion scale, which has been seen in a wide range of systems like solar wind, supernovae remnants, interstellar circumstances, galaxy evolutions, cluster merger events, and astrophysical laboratory studies. Turbulence generation assists the energy cascades from large to small scales, which eventually causes the energization or heating of the charged particles in laboratory and astrophysical plasmas [15–18]. The enhancement of the magnetic field and turbulence generation have been reported by the interaction of plasma flows generated by intense laser at the scale of laboratory astrophysics relevant to galaxy collisions, clumpy media, cluster merger events, etc [12, 13, 19]. The filamentary structure formation, magnetic field enhancement, turbulent spectra at electron and ion scale, acceleration of the charged particles, etc, have been observed in magnetized plasma generated by intense laser at the scale of laboratory astrophysics relevant to the interstellar medium, supernovae remnants, etc [11, 14, 20].

In a high-intensity laser-plasma interaction investigation conducted by Mondal *et al* [11], electron scale turbulent spectra with power-law scaling of $\sim k^{-2.0}$ (where k is the wavenumber) was recorded. The filamentary structure formations and magnetic field amplification of the order of mega gauss was also seen. This investigation is governed by electrons, and Wiebel instability is anticipated for this electron-dominated dynamics for early timescales up to the first few picoseconds. The intense laser-plasma interaction investigation

© 2024 IOP Publishing Ltd

performed by Chatterjee *et al* [20] reported the transition from the electron to the ion-dominated regions at later timescales. Also, two different spectral indices were observed in this investigation, separated by a spectral kink (the breakpoint at $k\rho_i \sim 1$, where k is the wave vector of the pump wave and ρ_i is the ion cyclotron length) in the turbulent spectra. The Weibel instability failed to explain the nonlinear dynamics and observations reported by Chatterjee *et al* [20] at later timescales due to the saturation of the Weibel instability. It was anticipated that after the saturation of Weibel instability, an unknown instability may be responsible for this nonlinear energy cascade at later timescales. Therefore, different mechanisms are needed to know about this unknown instability. A model based on the nonlinear wave-wave interaction considering ponderomotive and relativistic non linearity is proposed in this investigation, which could be one of the possible mechanisms.

Several investigations have attempted to replicate numerous space and astrophysical situations at the laboratory scale using the nonlinear wave-wave interaction model incorporating the ponderomotive effect [10, 21, 22]. The formation of filamentary structures, amplification of magnetic field, acceleration of charged particles, turbulent spectra at electron scale as well as at ion scales, and many other astrophysical situations have been observed within the magnetized plasma by utilizing the nonlinear interaction of the plasma waves model [23]. Singh *et al* [24] performed an investigation based on a nonlinear wave-wave coupling model and observed turbulent magnetic field amplification relevant to the cluster merger events. Singh *et al* [23] proposed a model based on the nonlinear wave-wave interaction between an elliptically polarized high-frequency pump wave and an electrostatic electron wave propagating perpendicular to the magnetic field to replicate the observations of electron-dominated dynamics performed by Mondal *et al* [11] in intense laser-plasma interaction experiments. In that investigation, enhancement of the magnetic field, scale size of the filamentary structures, and electron scale turbulence generation with the power-law scaling $\sim k^{-1.5}$ have been observed from the simulation findings. Some simulation results are quite close to the experimental results observed from the Mondal *et al* [11] intense laser-plasma interaction investigation. Still, there is a difference in the scaling of the turbulent spectra at the electron scale. Therefore, there is need for further investigation in this direction especially to look at electron-scale magnetic turbulent spectra. Hence, in this context, a model based on the nonlinear interaction of a circularly polarized pump wave with a parallel propagating electron plasma wave is proposed in this investigation.

This paper presents an investigation of a nonlinear wave-wave coupling of the right circularly polarized wave (R-mode wave) with the electron plasma wave (EPW) propagating parallel to the background magnetic field. The primary focus of this investigation is the electron scale turbulence generation associated with the coherent structure formation in the laser-produced magnetized plasma. In this context, model coupled equations for R-mode wave and EPW are developed, taking relativistic electron mass variation and nonlinear ponderomotive force into account. The equation of motion and all four of Maxwell's equations are employed to develop the model coupled equations. The laboratory simulations are used to solve the developed coupled equations utilizing pseudo-spectral and finite difference methods. This study also presents a simplified model considering the paraxial limit to look deep into the insights of the filamentary structure formation in the magnetized plasma.

The study is organized as follows. The model equations of R-mode wave and EPW are formulated in section 2; the normalized coupled model equations, details of the laboratory simulations, and used parameters are given in Section 3; the observed results from the numerical simulation and simplified model are discussed in Section 4; Results obtained from the simplified model are discussed in Section 5 and 6 provides the summary of this study.

2. Formulation of model equations

2.1. Basic model equations

The basic equations utilized for the mathematical formulation of nonlinear interaction between R-mode wave and EPW considering background magnetic field along z-direction ($\vec{B} = B_0\hat{z}$) are given in this section, which is as follows:

$$\nabla \times \vec{E} = -\frac{1}{c}\frac{\partial \vec{B}}{\partial t} \text{ (Faraday's law)} \tag{1}$$

$$\nabla \times \vec{B} = \frac{4\pi}{c}\vec{J} + \frac{1}{c}\frac{\partial \vec{E}}{\partial t} \text{ (Ampere's law)} \tag{2}$$

$$\nabla \cdot \vec{E} = -4\pi e n_e \text{ (Poisson's equation)} \tag{3}$$

$$\frac{\partial n_e}{\partial t} + \nabla \cdot (n_e \vec{v}_e) = 0 \text{ (Equation of continuity)} \tag{4}$$

$$\frac{\partial m_e \vec{v}_e}{\partial t} + m_e(\vec{v}_e \cdot \nabla)\vec{v}_e = -e\left(\vec{E} + \frac{(\vec{v}_e \times \vec{B})}{c}\right) \text{(Equation of motion)} \tag{5}$$

Where e, v_e, n_e, c and m_e are the electronic charge, the velocity of electrons, the number density of electrons, the speed of light in free space, and the electron mass, respectively.

2.2. Model equation for r-mode wave

Considering the laser beam propagating parallel to the background magnetic field ($\vec{B} = B_0\hat{z}$) with frequency ω_0 in the magnetized plasma. By combining equations (1) and (2), one may get the following wave equation:

$$\nabla^2\vec{E} - \nabla(\nabla \cdot \vec{E}) = \frac{1}{c^2}\frac{\partial^2\vec{D}}{\partial t^2} \tag{6}$$

Where $\vec{D}$ is the electric displacement vector and is defined as $\vec{D} = \bar{\bar{\varepsilon}} \cdot \vec{E}$, $\bar{\bar{\varepsilon}}$ denotes the plasma dielectric tensor [25] and $\vec{E}$ denotes electric field of the laser and it's varying rapidly. It is quite helpful to segregate the rapidly moving Electric field of the laser by writing $\vec{E}$ in the following way in order to adopt the slowly varying amplitude approximation [26, 27]:

$$\vec{E}(x, y, z, t) = E_0(x, y, z, t)e^{-i\omega_0 t} \tag{7}$$

Where the slowly varying amplitude of the laser pulse is denoted by E_0. Breaking equation (6) in x and y components, one can obtain the following equations:

$$\frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial y^2} - \frac{\partial}{\partial x}\left(\frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}\right) = -\frac{\omega_0^2}{c^2}(\bar{\bar{\varepsilon}} \cdot \vec{E})_x \tag{8}$$

$$\frac{\partial^2 E_y}{\partial x^2} + \frac{\partial^2 E_y}{\partial z^2} - \frac{\partial}{\partial y}\left(\frac{\partial E_x}{\partial x} + \frac{\partial E_z}{\partial z}\right) = -\frac{\omega_0^2}{c^2}(\bar{\bar{\varepsilon}} \cdot \vec{E})_y \tag{9}$$

Now, the component of equation (6) along the z-direction utilizing $\nabla \cdot \vec{D} = 0$ can be written as the following:

$$\frac{\partial E_z}{\partial z} \approx -\frac{1}{\varepsilon_{zz}}\left(\varepsilon_{xx}\frac{\partial E_x}{\partial x} + \varepsilon_{xy}\frac{\partial E_y}{\partial x} + \varepsilon_{yx}\frac{\partial E_x}{\partial y} + \varepsilon_{yy}\frac{\partial E_y}{\partial y}\right) \tag{10}$$

Where $\varepsilon_{zz} = 1 - (\omega_{pe}^2/\omega_0^2)$ It is assumed that $E_+ = E_x + iE_y$ and $E_- = E_x - iE_y$ represents two modes of propagation of pump wave, namely L-mode (left circularly polarized) and R-mode (right circularly polarized) respectively. For the R-mode wave, the electric field vector rotates in a clockwise direction, and for the L-mode wave, the electric field vector rotates in a counter-clockwise direction. Now, multiplying the equation (9) by $\pm i$ and adding it to equation (8) and then Substituting the value of $\partial E_z/\partial z$ from equation (10), the equations for the two coupled modes (R-mode and L-mode) can be written as follows:

$$\begin{aligned}&\frac{\partial^2 E_+}{\partial z^2} + \frac{1}{2}\left(1 + \frac{\varepsilon_+}{\varepsilon_{zz}}\right)\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)E_+ + \frac{1}{2}\left(-1 + \frac{\varepsilon_-}{\varepsilon_{zz}}\right)\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}\right)E_- \\ &- i\left(1 - \frac{\varepsilon_-}{\varepsilon_{zz}}\right)\frac{\partial^2 E_-}{\partial x \partial y} + \frac{\omega_0^2}{c^2}(\varepsilon_+)E_- = 0\end{aligned} \tag{11}$$

$$\begin{aligned}&\frac{\partial^2 E_-}{\partial z^2} + \frac{1}{2}\left(1 + \frac{\varepsilon_-}{\varepsilon_{zz}}\right)\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)E_- + \frac{1}{2}\left(-1 + \frac{\varepsilon_+}{\varepsilon_{zz}}\right)\left(\frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2}\right)E_+ \\ &+ i\left(1 - \frac{\varepsilon_+}{\varepsilon_{zz}}\right)\frac{\partial^2 E_+}{\partial x \partial y} + \frac{\omega_0^2}{c^2}(\varepsilon_-)E_- = 0\end{aligned} \tag{12}$$

Where $\varepsilon_- = \varepsilon_{xx} + i\varepsilon_{xy} = 1 - \frac{\omega_{pe}^2/\omega_0^2}{1 - \omega_{ce}/\omega_0}$, $\varepsilon_+ = \varepsilon_{xx} - i\varepsilon_{xy} = 1 - \frac{\omega_{pe}^2/\omega_0^2}{1 + \omega_{ce}/\omega_0}$. E_+ and E_- are two loosely coupled modes. It is challenging to solve the simultaneous propagation of E_+ and E_- hence assuming E_+ to be zero [27] (by considering $E_- = 0$, an analogous equation can also be derived for E_+), the equation for R-mode wave can be written as:

$$\frac{\partial^2}{\partial z^2} + \delta_-\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)E_- + \frac{\omega_0^2}{c^2}(\varepsilon_-)E_- = 0 \tag{13}$$

Here, $\delta_- = 1/2(1 + (\varepsilon_-/\varepsilon_{zz}))$, $\omega_{pe}^2 = 4\pi n_e e^2/m_e$ (here $n_e = n_0 + n_1$, where n_0 is the unperturbed background electron number density and n_1 is the perturbed electron number density) is the plasma frequency and $\omega_{ce} = \frac{eB_0}{m_e c}$ is the electron cyclotron frequency. Now the solution of equation (13) can be written as (Because there is symmetry in x and y direction hence for simplification here only one direction is considered)

$$E_- = E_{0-}(x,\ z,\ t)e^{(-i(\omega_0 t - k_- z))} \tag{14}$$

Here $k_- = \omega_0/c(\varepsilon_-)^{\frac{1}{2}}$ is the propagation constant of the R-mode wave in the plasma having a background magnetic field. Now, substituting equation (14) in equation (13) and then applying slowly varying envelop approximation $\left(\frac{\partial^2 E_{0-}}{\partial t^2} \ll \omega_0 \frac{\partial E_{0-}}{\partial t}\right)$ and using the relativistic mass variation effect, we get the following equation:

$$2ik_-\frac{\partial E_{0-}}{\partial z} + \frac{\partial^2 E_{0-}}{\partial z^2} - k_-^2 E_{0-} + \delta_-\frac{\partial^2 E_{0-}}{\partial x^2} + 2i\omega_0\frac{\varepsilon_-}{c^2}\frac{\partial E_{0-}}{\partial t} + \frac{1}{c^2}\left(1 - \frac{\frac{\omega_p^2}{\gamma\omega_0^2}}{1 - \frac{\omega_{ce}}{\gamma\omega_0}}\right)\omega_0^2 E_{0-} = 0 \tag{15}$$

Where γ is the Lorentz factor and is given as $\gamma = (1 - v_0^2/c^2)^{-1/2}$, here $v_0^2 = v_{0x}^2 + v_{0y}^2$ where v_{0x} and v_{oy} are the velocity components of the pump wave and are calculated by using the equation of motion (equation (5)) and are given as:

$$v_{0x} = \frac{e}{m_e\omega_0}\left(-iE_x - \frac{\omega_{ce}}{\omega_0}E_y\right)\left(1 - \frac{\omega_{ce}^2}{\omega_0^2}\right)^{-1} \tag{16}$$

$$v_{0y} = \frac{e}{m_e\omega_0}\left(-iE_y + \frac{\omega_{ce}}{\omega_0}E_x\right)\left(1 - \frac{\omega_{ce}^2}{\omega_0^2}\right)^{-1} \tag{17}$$

Using these velocity components γ comes out to be $1 + \alpha\ |E_{0-}|^2$ where $|E_{0-}|^2 = E_x^2 + E_y^2$ and $\alpha = \frac{1}{2c^2}\left(\frac{e}{m_e\omega_0}\right)^2\left[\frac{(1+\omega_{ce}^2/\omega_0^2)}{(1-\omega_{ce}^2/\omega_0^2)^2}\right]$. By utilizing the Lorentz factor ($\gamma = 1 + \alpha\ |E_{0-}|^2$) and density perturbation ($n_e = n_0 + n_1$) in equation (15), one can easily get the following expression:

$$\left(2ik_-\frac{\partial}{\partial z} + \frac{\partial^2}{\partial z^2} + \delta_-\frac{\partial^2}{\partial x^2} + 2i\omega_0\frac{\varepsilon_-}{c^2}\frac{\partial}{\partial t}\right)E_{0-} = \left(\frac{\omega_{pe}^2}{c^2(1 - \omega_{ce}/\omega_0)}\left(\frac{n_1}{n_0} - \alpha\ |E_{0-}|^2\right)\right)E_{0-} \tag{18}$$

The above equation (18) represents the dynamical coupled equation for the pump right-circularly polarized (R-mode) wave.

2.3. Model equation for electron plasma wave (EPW)

In magnetized plasma, considering propagation of the EPW parallel to the laser propagation direction. EPW is the electrostatic wave and its electric field vector in the propagation direction; hence, the electric field expression of the electron plasma wave can be written as follows:

$$\vec{E}_1 = E_{1z}\exp\left(i(k_1 z - \omega_1 t)\right)\hat{z} \tag{19}$$

where ω_1 and k_1 are the oscillation frequency and the propagation constant for the EPW, respectively. Now, considering the thermal effects and ponderomotive force equation (5) can be written as:

$$m_e\frac{\partial\vec{v}_{1e}}{\partial t} = -e\vec{E}_1 - \frac{e}{c}(\vec{v}_{1e} \times B_0\hat{z}) - \frac{K_B T_e \nabla n_1}{n_0} + \vec{F}_0 \tag{20}$$

Where $\vec{v}_{1e}$ is the velocity of electrons in the electron plasma wave and $\vec{F}_0$ is the ponderomotive force that is exerted by the pump wave on the electron plasma wave inside the plasma and the expression for the ponderomotive force is given as:-

$$\vec{F}_0 = -\frac{1}{2}\left\{m_e(\vec{v}_0 \cdot \nabla)\vec{v}_0 + \frac{e}{c}(\vec{v}_0 \times \vec{B}_p)\right\} \tag{21}$$

Here, $\vec{B}_p$ is the magnetic field of the pump wave whose propagation direction can be decided using Faraday's law (equation (1). Taking the time derivative of the equation of continuity (equation (4)) and applying the procedure of linearization [28] we get the following equation:

$$\frac{\partial^2 n_1}{\partial t^2} + n_0\nabla \cdot \frac{\partial\vec{v}_{1e}}{\partial t} = 0 \tag{22}$$

Calculating $\partial\vec{v}_{1e}/\partial t$ from equation (20), substituting it in equation (22), and taking the z-component of the obtained equation, we get the following equation:

$$\frac{\partial^2 n_1}{\partial t} - \frac{n_0 e\nabla_z E_{1z}}{m_e} - \frac{n_0}{m_e}\frac{K_B T_e}{n_0}\frac{\partial^2 n_1}{\partial z^2} + \frac{n_0 F_{0z}}{m_e} = 0 \tag{23}$$

The z-component of the ponderomotive force can be calculated by equation (21), and using the velocity components of the pump wave given by equations (16) and (17), the expression for the ponderomotive force is given as:

$$F_{0z} = \frac{k_0 e^2 |E_{0-}|^2}{2i\omega_0^2 m_e (1 - \omega_{ce}^2/\omega_0^2)} \tag{24}$$

Using Poisson's equation (equation (3)) and equation (24) in equation (23) and rearranging the obtained equation, we get the following equation:

$$\left(\frac{\partial^2}{\partial t^2} + \omega_{pe}^2 - v_{th}^2 \frac{\partial^2}{\partial z^2}\right) n_1 = \left(\frac{\omega_{pe}^2}{8\pi m_e(\omega_0^2 - \omega_{ce}^2)} \frac{\partial^2}{\partial z^2}\right) |E_{0-}|^2 \tag{25}$$

The above equation (25) represents the model coupled equation for electrostatic electron plasma wave (EPW).

3. Normalised equations and parameters used

Numerical simulation has been used in this section to investigate the dynamics associated with the interaction of R-mode waves and Plasma waves. The coupled system of equations (18) and (25) in the dimensionless form can be expressed as:-

$$\left(i\frac{\partial}{\partial t'} + ic_1 \frac{\partial}{\partial z'} + \frac{\partial^2}{\partial z'^2} + \frac{\partial^2}{\partial x'^2}\right) E_{0-}{}' = (c_4 n'_1 - c_2 |E_{0-}{}'|^2) E_{0-}{}' \tag{26}$$

$$\left(i\frac{\partial^2}{\partial t'^2} + c_6 - c_3 \frac{\partial^2}{\partial z'^2}\right) n'_1 = c_5 \frac{\partial^2}{\partial z'^2} |E_{0-}{}'|^2 \tag{27}$$

Where $E_{0-}{}'$, $n_1{}'$ is normalized using E_n and n_n respectively and $c_1 = \frac{k_- t_n c^2}{\omega_0 \varepsilon_- z_n}$, $c_2 = \frac{\alpha c^2 t_n E_n^2}{2\omega_0 \varepsilon_-}$, $c_3 = \frac{v_{th}^2 t_n^2}{z_n^2}$, are the normalizing constants and the normalizing quantities are $x_n = \sqrt{\frac{\delta_- c^2 t_n}{2\omega_0 \varepsilon_-}}$, $z_n = \sqrt{\frac{c^2 t_n}{2\omega_0 \varepsilon_-}}$ $t_n = 5\omega_{pe}^{-1}$, $E_n = \sqrt{\frac{8\pi m_e(\omega_0^2 - \omega_{ce}^2) z_n^2 n_n}{t_n^2 \omega_{pe}^2}}$, $v_{th} = \sqrt{\frac{K_B T_e}{m_e}}$. The coupled equations represented by equations (28) and (29) are solved for the electric field, and then equation (1) can be utilized to transform the electric field into the magnetic field. With a step size of 10^{-5}, the finite difference technique for temporal integration is used along with the predictor-corrector approach, and for the space integration, a pseudo-spectral scheme in the x-z plane has been employed. The coupled system of dimensionless equations (26) and (27) are solved by computational simulations in this study. A well-known nonlinear Schrodinger equation (NLS) is used to check the accuracy of the algorithm that we have used for our computation simulations. Up to the accuracy of 10^{-5}, the Plasmon number was preserved during the simulation. The algorithm was altered to solve the dimensionless nonlinear coupled equations. The following initial conditions are used for the computational simulation:

$$E_{m-}(x, z, 0) = |E_{m0-}| \{1 + p \cos(\alpha_x x)\} \{1 + p \cos(\alpha_z z)\} \tag{28}$$

$$n_1(x, z) = \sum_m |E_{m-}(x, z, 0)|^2 \tag{29}$$

Where E_{m0-} and p denote the initial amplitude of the laser and the perturbation amplitude, respectively. Numerical simulations using the above given initial conditions are implemented to solve the coupled system of equations (26) and (27) in the 2-dimensional periodic spatial domain. $\frac{2\pi}{\alpha_x} \times \frac{2\pi}{\alpha_z}$ is the size of the periodic spatial domain with 512×512 grid points, and $\alpha_x = \alpha_z = 0.1$ is the wavenumbers of perturbation that are normalized by x_n^{-1}, z_n^{-1} respectively. The parameters used for the numerical simulation are $T_e = 8$ keV, $B_0 = 0.1$ MG, $n_0 = 5.75 \times 10^{17}$ cm^{-3}, $\omega_0 = 2.25 \times 10^{14}$ rad/s, $\omega_{ce} = 1.75 \times 10^{12}$ rad/s, $\omega_{pe} = 4.27 \times 10^{13}$ rad/s and the values of the normalizing parameters are $n_n \approx 0.01 n_0$, $t_n = 0.12ps = 5\omega_{pe}^{-1}$, $x_n \approx 5\mu m \approx 0.7\lambda_e$, $z_n \approx 5\mu m \approx 0.7\lambda_e$(where $\lambda_e = \frac{c}{\omega_{pe}}$ is the electron skin depth) and $E_n = 2.85 \times 10^5$ statV/cm. Condition for normalization is such that the time is normalized with plasma frequency and space coordinates are normalized with electron skin depth. This study aims to understand the physics behind the magnetic turbulence generation at the electron scale associated with the localized structure formation reported by Mondal *et al* [11] in experimental investigations of intense laser-plasma interaction. In this context, numerical simulation is performed in the present investigation utilizing the parameters from the experimental investigation, considering our simulation constraints.

4. Results and discussion

The numerical simulations are carried out to solve the dimensionless coupled equations (26) and (27), and the observed results are discussed in this section. A ponderomotive force is exerted on the plasma wave during the propagation of the R-mode wave inside the plasma medium that causes perturbations in the background

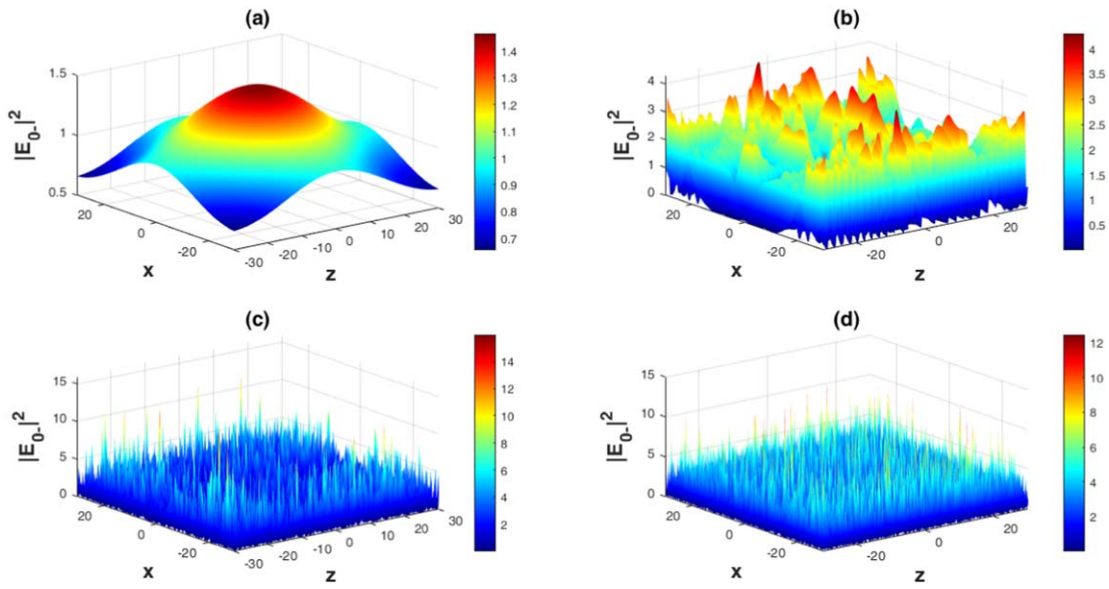


Figure 1. For different normalized times (a) 0, (b) 15, (c) 20, and (d) 30, the Normalized intensity of the R-mode wave in the x-z plane.

density, resulting in the emergence of density cavities and humps. When the R-mode wave further propagates in the modified background density, breaking of the R-mode wave in the filamentary (localized) structures takes place, and these structures become chaotic with time evolution, indicating the turbulence generation. It has been observed from the simulation results that the initiation time of the filamentary structures is approximately 1.1 picoseconds, and these structures become localized at around 1.7 picoseconds.

The localization initiating time in the experimental results reported by Mondal *et al* [11] was mentioned around 0.3 picosecond, which is quite close to our simulation observations. The localization profiles of the R-mode wave at different normalized times in the *x*-*z* plane are illustrated in figure 1. Figures 1(a) and (b) show the initial profile (at normalized time t = 0) of the pump wave and early stage of the localized structures (at normalized time t = 15) respectively inside the plasma medium. Figures 1(c)–(d) illustrate the chaotic localized structures at normalized times t = 20 and t = 30 of the R-mode wave inside the plasma medium, indicating the turbulence generation. The energy exchange between the R-mode wave and the plasma wave is shown by the evolution of the localized structures.

Figures 2(a)–(b) shows density variation in an *x*-*z* plane at a normalized time t = 24 and t = 30, respectively. This depiction portrays the humps and depletion of the density in various locations in the x-z plane.

Figure 3 complements the understanding by presenting the localization plots of the field and density cavitation. Mondal *et al* [11] show a filament size of ∼4 μm, and the result obtained in this study shows a thickness of about ∼6 μm at an early time, which is close to the results obtained by Mondal *et al* [11]. The resolution of the system is∼ 0.6 μm, and hence, the results obtained in this paper are adequate to resolve the filamentary structures satisfactorily. Thickness of the localized structures at the later time is around ∼2 μm, which is also resolvable by the system. The localization of the pump wave takes place inside the magnetic plasma medium, and the pump energy transports to the small scale from the large scale, which can be indicated by the power spectrum. For the long spatial scale length, the magnetic field power spectrum (electron dominated) of the R-mode wave versus *k* has been plotted. Figure 4. shows the ensemble-averaged power spectrum, which is averaged over the normalized time from t = 20 and t = 30 comprising ten spectra. To confirm the turbulent scaling, a reference red solid line is given in the power spectrum of the order of $\sim k^{-1.9}$. This power law scaling is quite close to that observed by Mondal *et al* [11] ($\sim k^{-2.0}$). In the magnetic field profile, the emergence of chaotic structures is an indication of turbulence generation.

To understand the evolution of the density harmonics/modes for different points in time, the fluctuation of the modes of density $|n_m|$ versus k_x (when $k_z = 0$) has been studied. At different times, the fluctuation of the density modes with k_x is shown in figures 5(a)–(b). At $k_x = \alpha_x, 2\alpha_x, 3\alpha_x$, up to $256\alpha_x$ the spectral density modes are generated, and these modes exhibit variation over time. It can be made clear from equation (27) that the nonlinear ponderomotive force {last term of equation (27)} is the cause of the origin of density harmonics for the different instants of time.For $k_z = 0$, at a given normalized time (t), the generation of the modes of density is provided by $n_1'(x, z) = \{1 + p \cos(\alpha_x x)\}^2$

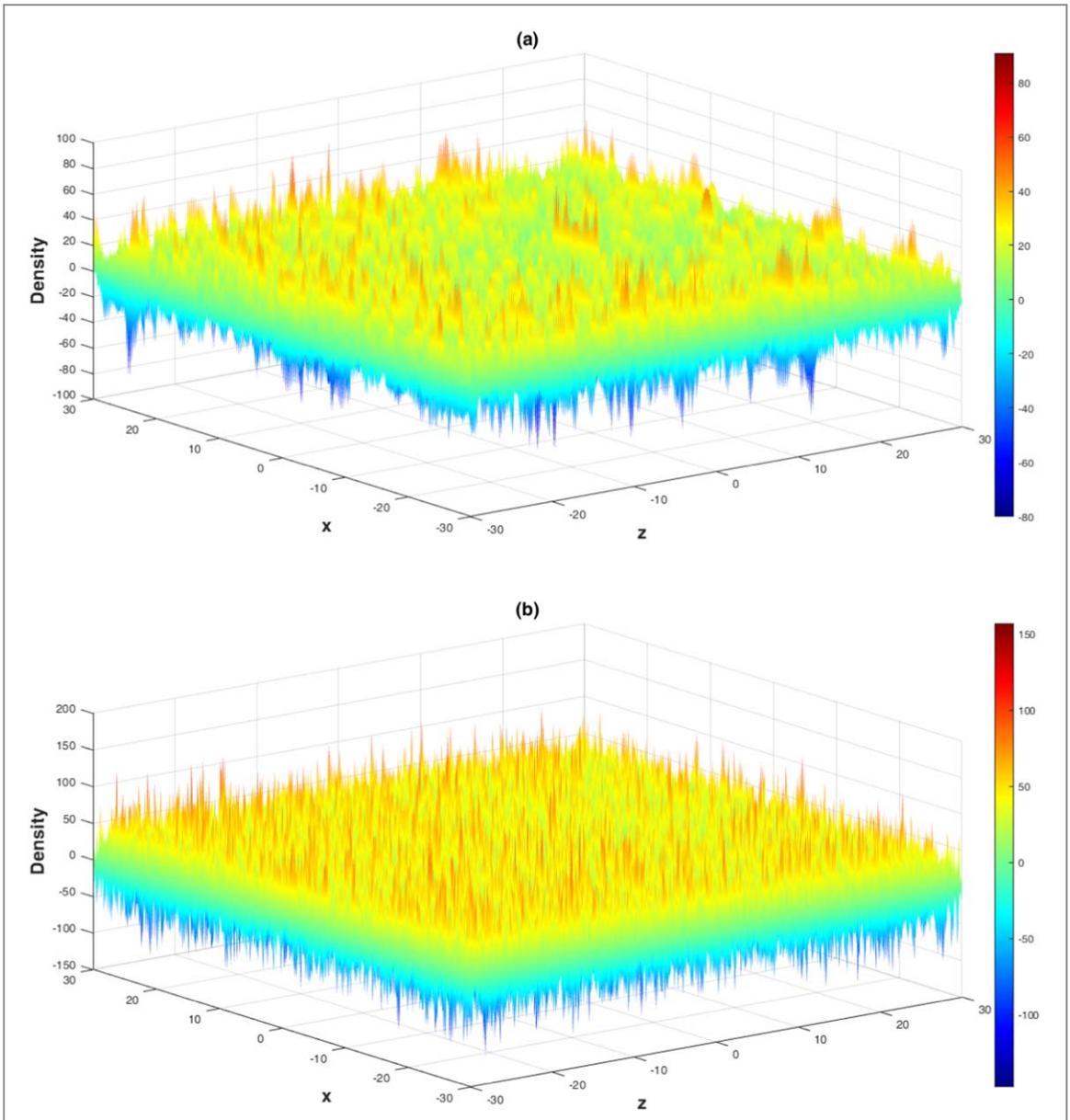


Figure 2. For normalized time (a) t = 24 and (b) t = 30 profile of the normalized density in the x-z plane.

$$n_1'(x, z) = \{1 + p^2 \cos(\alpha_x x)^2 + 2p \cos(\alpha_x x)\} \tag{30}$$

Equation (30) shows the generation of mode occurs at α_x and 2 α_x. Using equation (30), the coupling term $n_1'E_{0-}'$ of equation (26) can be expressed as

$$n_1'E_{0-}' = \{1 + p^2 \cos(\alpha_x x)^2 + 2p \cos(\alpha_x x)\}\{1 + p \cos(\alpha_x x)\} \tag{31}$$

The higher modes will be generated in E_{0-}' at α_x, $2\alpha_x$, and $3\alpha_x$ it can be seen from equation (31). The phenomenon of generation of higher density modes will persist until k_x becomes equal to $256\alpha_x$ in the spectrum.

5. Simplified model

A semi-analytical approach is also provided in this investigation to explore the evolution of pump wave. However, turbulence generation observation is not possible in this model due to the consideration of paraxial limits. Still, the nonlinear evolution of the pump wave inside the magnetized plasma is also an important phenomenon. Therefore, the exact comparison of the results obtained by this model and simulation results is not possible.

After neglecting the time variation from the equation (18), the equation for the R-mode wave can be written as follows:

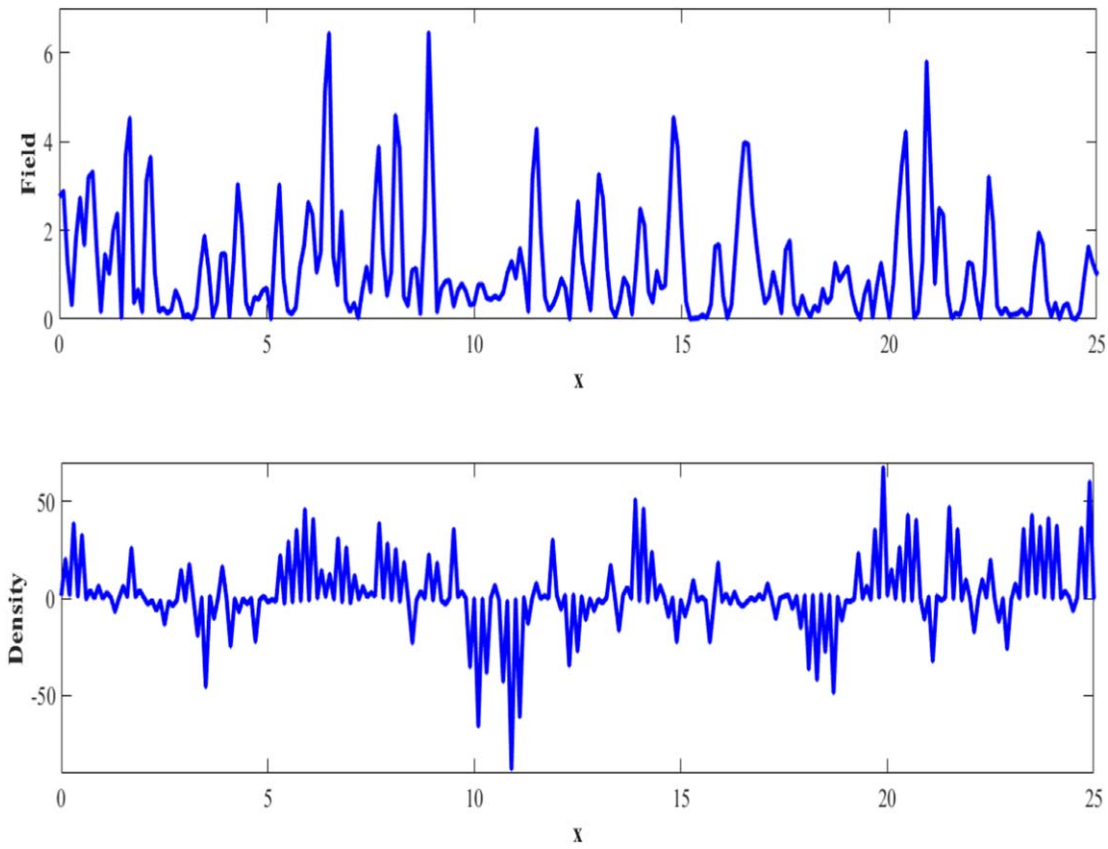


Figure 3. At normalized time t = 26, the profile of localization of field and density cavitation for R-mode wave.

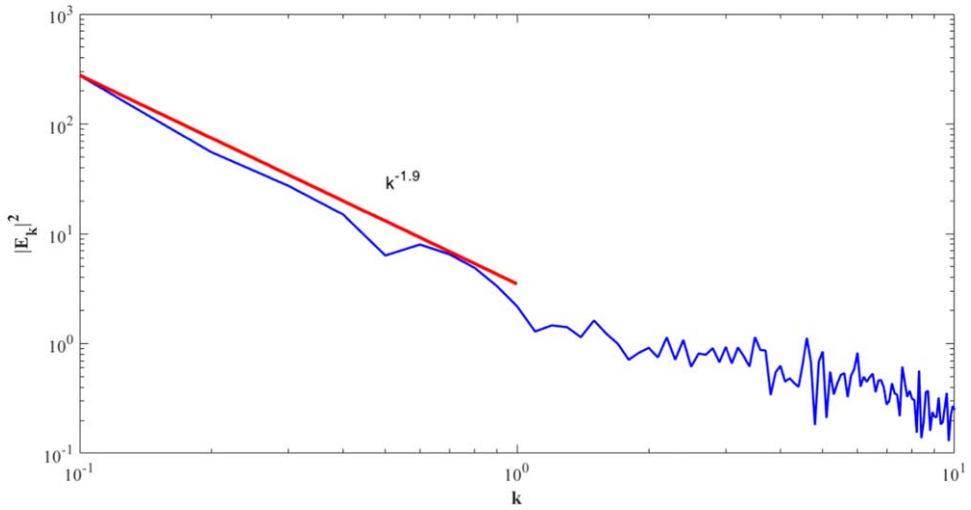


Figure 4. The magnetic power spectrum that is ensemble-averaged over the normalized time between t = 20 and t = 30.

$$2ik_{-}\frac{\partial E_{0-}}{\partial z}+\delta_{-}\frac{\partial^2 E_{0-}}{\partial x^2}+2i\omega_0\frac{\varepsilon_{-}}{c^2}\frac{\partial E_{0-}}{\partial t}-\phi\left(\frac{n_1}{n_0}-\alpha\left|E_{0-}\right|^2\right)E_{0-}=0 \tag{32}$$

Here $\phi=\frac{\omega_{pe}^2}{c^2(1-\omega_{ce}/\omega_0)}$ and the term (n_1/n_0) is time-dependent and can be written as:

$$\frac{n_1}{n_0}=N+\sum_{m=1}^{256}n_m\{\cos(m\alpha_x x)\}=N+\sum_{m=1}^{256}n_m\left(1-\frac{(m\alpha_x x)^2}{2}+\ldots\right) \tag{33}$$

n_m Represents number density modes evaluated by simulations for $m=1-256$, N denotes the Plasmon number in equation (33), and. Now, using the eikonal, the solution of equation (33) is assumed as follows [29]:

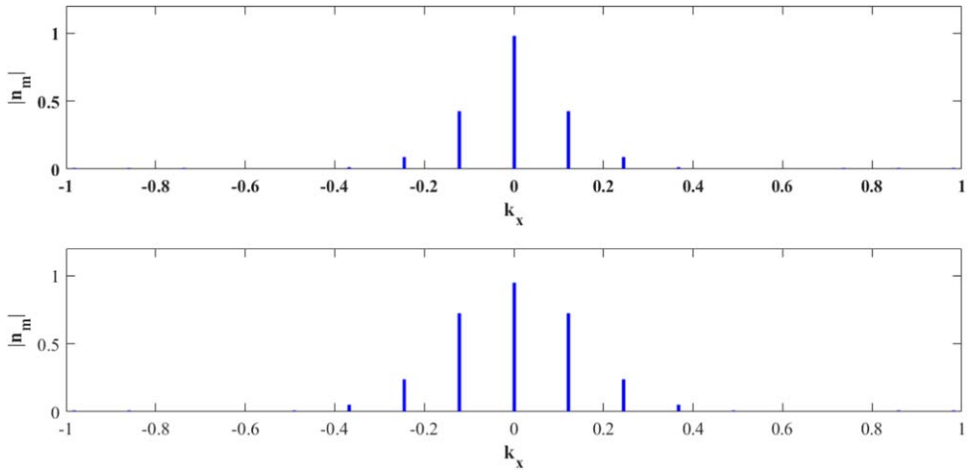


Figure 5. At normalized times (a) 10 & (b) 15 and at fixed k_z, fluctuation of density ($|n_m|$) versus k_x.

$$E_{0-} = A_0(x, z)\exp(ik_s(x, z)) \tag{34}$$

Now, substituting equations (34) and (33) in equation (32) and isolating the imaginary and real components, the following equations are obtained:

$$2k_^2 A_0\frac{\partial s}{\partial z} - \delta_\frac{\partial^2 A_0}{\partial x^2} + k_^2\delta_A_0\left(\frac{\partial s}{\partial x}\right)^2 + \phi\left\{N + \sum_{m=1}^{256} n_m\left(1 - \frac{(m\alpha_x x)^2}{2} + ...\right) - \alpha\,|A_0|^2\right\}A_0 = 0 \tag{35}$$

$$2\frac{\partial A_0}{\partial z} + 2\delta_\frac{\partial A_0}{\partial x}\frac{\partial s}{\partial x} + \delta_A_0\frac{\partial^2 s}{\partial x^2} = 0 \tag{36}$$

Now, with the initial beam width r_0 defining the amplitude of the beam and eikonal as:

$$A_0 = \frac{E_{00}}{\sqrt{f}}\exp\left(-\frac{x^2}{2r_0^2 f^2}\right) \tag{37}$$

And

$$s = \frac{x^2}{2f\delta_}\frac{\partial f}{\partial z} \tag{38}$$

Here, r_0, f, and E_{00} denote the transverse scale size, beam width parameter which is dimensionless, and initial field of the laser beam, respectively. The dynamical equation for the beam width parameter can be obtained by putting equations (37) and (38) in equation (36) and comparing the x^2 coefficients, which is given by the following equations:

$$\frac{d^2 f}{d\xi^2} = \frac{\delta_^2}{f^3\varepsilon_} - \frac{\phi E_{00}^2\delta_r_0^2}{E_n^2 f^2\varepsilon_} + \frac{\phi f\delta_r_0^4\alpha_x^2}{2\varepsilon_}\sum_{m=1}^{256} n_m m^2 \tag{39}$$

Here $\xi\left(=\frac{c}{\omega_0 r_0^2}\right)z$. Equation (39) describe the beam width parameter's behaviour as the laser travel through the plasma. In equation (39) the natural diffraction term is represented by the first term this term is responsible for the divergence of laser beam, the nonlinear term due to ponderomotive force and nonlinear term due to variations in background density with time is represented by second term and the third term respectively and the second and the third terms are responsible for the focusing of laser beam inside plasma. To solve equation (39) with numerical simulation parameters, the Runge–Kutta method of fourth order is utilized using initial conditions $f|_{x=0}$ and $\frac{df}{d\xi}\Big|_{x=0}$. The transient development of the R-mode wave is attributed to the inclusion of the time-varying background density variations in the third term of equation (39). Wave can undergo without convergence or divergence in the plasma when there is a balance between the nonlinear term (second and third terms) and the diffraction term (first term). It can be made clear from equation (39) that the self-focusing and defocusing of the laser beam is due to the competition between the diffraction term and the nonlinear term. Hence, one can say that filamentary structures are formed due to the nonlinearity exerted by the R-mode wave on the plasma wave. Self-focusing of the laser beam inside the plasma occurs when the nonlinear terms predominate the diffraction term, and at this point, the beam width parameter gets its minimum value. At the minimum value of beam width parameter (f), the intensity of the laser beam becomes maximum inside the plasma medium. Figure 6 illustrates the profile of the normalized intensity of the R-mode wave versus the

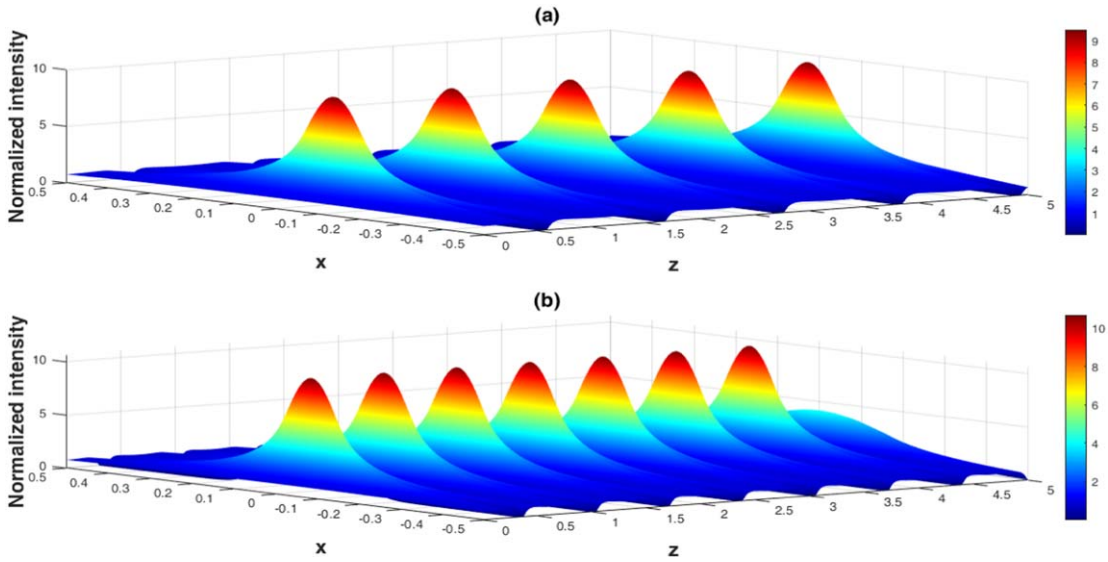


Figure 6. At various normalized times (a) t = 10 (b) t = 15, normalized intensity profiles.

dimensionless propagation distance ξ at two different normalized times. It can be noticed from figure 6 that the variations in the normalized intensity of the R-mode wave inside the plasma medium occur over time with density harmonics variations. This model helps in understanding the localization process, which is the main motive behind semi-analytical study.

6. Summary and conclusion

Numerical techniques and computer simulations are used to solve the nonlinear coupled system of R-mode wave (pump wave) and electron plasma wave. The ponderomotive and relativistic nonlinearities are responsible for the nonlinear coupling of the R-mode and EPW, as result the nonlinear evolution of the R-mode wave have been observed inside the plasma medium. As time proceeds, the pump wave localization becomes more chaotic leading to the turbulence generation. Moreover, the density modifications in the form of cavitation and humps have been observed, which corresponds to the electron plasma waves. The generation of magnetic turbulence associated with the localization of the R-mode wave inside the plasma medium and energy cascading in plasma can be inferred. From the numerical simulations, the ensemble-averaged power spectrum was also obtained for this investigation with the power-law scaling $\sim k^{-1.9}$. The some of the simulation findings observed from the present investigation, like nonlinear evolution (filamentation) of the pump wave, initiation time of the localized structures and their typical scale sizes, and scaling of the turbulent spectra at electron scale quite resemble the experimental outcomes reported by Mondal *et al* [11] experimental investigations of intense laser-plasma interaction. Thus, the model based on the nonlinear interaction of plasma modes could be essential to get insight into the physics of the magnetized plasma associated with the laboratory, space, and astrophysical scenarios.

Acknowledgments

This work received partial support from the Department of Science and Technology (DST, Grant No.-CRG/2022/000546) and also one of the authors (Aditya) is thankful to the Delhi Technological University (DTU), New Delhi for providing fellowship grant to carry out Ph.D. work.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

ORCID iDs

Suresh C Sharma https://orcid.org/0000-0003-2170-195X
Indraj Singh https://orcid.org/0000-0001-7247-0747

References

[1] Remington B A, Drake R P and Ryutov D D 2006 Experimental astrophysics with high power lasers and Z pinches *Rev. Mod. Phys.* **78** 755–807
[2] Gregori G, Reville B and Miniati F 2015 The generation and amplification of intergalactic magnetic fields in analogue laboratory experiments with high power lasers *Phys. Rep.* **601** 1–34
[3] Wilson T C, Sheng Z M, Eliasson B and McKenna P 2021 Magnetic field amplification by high power lasers in underdense plasma *Plasma Phys. Control. Fusion* **63** 084001 84001
[4] Wagner R, Chen S-Y, Maksimchuk A and Umstadter D 1997 Electron acceleration by a laser wakefield in a relativistically self-guided channel *Phys. Rev. Lett.* **78** 3125
[5] Amiranoff F, Baton S, Bernard D, Cros B, Descamps D, Dorchies F, Jacquet F, Malka V, Marques J R and Matthieussent G 1998 Observation of laser wakefield acceleration of electrons *Phys. Rev. Lett.* **81** 995
[6] Klein U and Fletcher A 2014 *Galactic and Intergalactic Magnetic Fields* (Springer)
[7] Zweibel E G and Heiles C 1997 Magnetic fields in galaxies and beyond *Nature* **385** 131–6
[8] Bell A R 2004 Turbulent amplification of magnetic field and diffusive shock acceleration of cosmic rays *Mon. Not. R. Astron. Soc.* **353** 550–8
[9] Ryu D, Kang H, Cho J and Das S 2008 Turbulence and magnetic fields in the large-scale structure of the universe *Science (80-.).* **320** 909–12
[10] Singh I, Uma R and Sharma R P 2022 A nonlinear wave-based model to study the magnetic turbulence generation associated with the nonlinear evolution of Kinetic Alfven Wave in laser-produced plasma relevant to laboratory and astrophysical plasmas *Optik (Stuttg).* **267** 169698 169698
[11] Mondal S, Narayanan V, Ding W J, Lad A D, Hao B, Ahmad S, Wang W M, Sheng Z M, Sengupta S and Kaw P 2012 Direct observation of turbulent magnetic fields in hot, dense laser produced plasmas *Proc. Natl Acad. Sci.* **109** 8011–5
[12] Meinecke J, Doyle H W, Miniati F, Bell A R, Bingham R, Crowston R, Drake R P, Fatenejad M, Koenig M and Kuramitsu Y 2014 Turbulent amplification of magnetic fields in laboratory laser-produced shock waves *Nat. Phys.* **10** 520–4
[13] Meinecke J, Tzeferacos P, Bell A, Bingham R, Clarke R, Churazov E, Crowston R, Doyle H, Drake R P and Heathcote R 2015 Developed turbulence and nonlinear amplification of magnetic fields in laboratory and astrophysical plasmas *Proc. Natl Acad. Sci.* **112** 8211–5
[14] Fiuza F, Swadling G F, Grassi A, Rinderknecht H G, Higginson D P, Ryutov D D, Bruulsema C, Drake R P, Funk S and Glenzer S 2020 Electron acceleration in laboratory-produced turbulent collisionless shocks *Nat. Phys.* **16** 916–20
[15] Patnaude D J and Fesen R A 2009 Proper motions and brightness variations of nonthermal x-ray filaments in the Cassiopeia a supernova remnant *Astrophys. J.* **697** 535
[16] Barnes M, Parra F I and Dorland W 2012 Turbulent transport and heating of trace heavy ions in hot magnetized plasmas *Phys. Rev. Lett.* **109** 185003 185003
[17] Reynolds S P, Gaensler B M and Bocchino F 2012 Magnetic fields in supernova remnants and pulsar-wind nebulae *Space Sci. Rev.* **166** 231–61
[18] Bell A R and Lucek S G 2001 Cosmic ray acceleration to very high energy through the non-linear amplification by cosmic rays of the seed magnetic field *Mon. Not. R. Astron. Soc.* **321** 433–8
[19] Tzeferacos P *et al* 2018 Laboratory evidence of dynamo amplification of magnetic fields in a turbulent plasma *Nat. Commun.* **9** 591
[20] Chatterjee G, Schoeffler K M, Kumar Singh P, Adak A, Lad A D, Sengupta S, Kaw P, Silva L O, Das A and Kumar G R 2017 Magnetic turbulence in a table-top laser-plasma relevant to astrophysical scenarios *Nat. Commun.* **8** 15970 15970
[21] Singh I, Uma R and Sharma R P 2023 A simulation study of electron-scale turbulence generation by nonlinear wave–wave interaction in laser-produced plasmas *Phys. Plasmas* **30** 092107
[22] Singh I, Uma R and Sharma R P 2023 A fermi acceleration model to study the electron acceleration in laser-produced plasma with localized structures formation and turbulence generation *Phys. Plasmas* **30** 022108
[23] Singh I, Dewan H, Uma R and Sharma R P 2021 Nonlinear evolution of localized structures and turbulent magnetic field amplification by extraordinary laser beam in laser produced plasmas *Phys. Plasmas* **28** 092109
[24] Singh I, Uma R and Sharma R P 2023 Simulation to study the turbulent amplification of magnetic field relevant to laboratory and astrophysical plasmas *Eur. Phys. J. Plus* **138** 590
[25] Stix T H 1992 *Waves in Plasmas* (Springer Science & Business Media)
[26] Agrawal G P 2000 Nonlinear fiber optics *Nonlinear Science at the Dawn of the 21st Century* (Springer) 195–211
[27] Kumar S, Chauhan P K, Sharma R P and Uma R 2021 Compression of the laser pulse in magnetized plasma having relativistic regime *Optik (Stuttg).* **242** 167130 167130
[28] Chen F F 1984 *Introduction to Plasma Physics and Controlled Fusion* 1 (Springer)
[29] Akhmanov S A, Sukhorukov A P and Khokhlov R V 1968 Self-focusing and diffraction of light in a nonlinear medium *Sov. Phys. USPEKHI* **10** 609

Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

#LetsTalk: Understanding Social Media Usage of Self-Harm Users in India

Garima Chhikara[1,2], **Abhijnan Chakraborty**[1]

[1] Indian Institute of Technology Delhi, India
[2] Delhi Technological University, India

Abstract

Mental health concerns, such as depression, pose significant challenges for support systems in effectively identifying affected individuals. On the other hand, people suffering from depression often find it easier to discuss on social media rather than in face-to-face interactions. Additionally, the development of distressing conditions typically arises from a multitude of factors accumulated over time rather than a singular event. To gain a fine-grained understanding of these facets, in this work, we perform a longitudinal analysis of the tweeting behaviour of Indian users who post content related to self-harm. We categorise users based on their posting frequency and examine various aspects including their social network, bio descriptions, tweeting preferences, temporal variations and cognitive indicators. By elucidating these nuances, we aim to contribute insights that could aid in the early detection of mental health issues and prompt timely intervention from support networks.

Introduction

Today, a large number of people around the world suffer from mental health issues like depression and anxiety. Mental health support systems are in place to help people who are struggling with mental health problems. However, these systems often face the challenge of identifying people who need immediate help. This is because people who are depressed may not always seek help, or they may not be able to identify that they need help. Worryingly, unless attended to in early stages, many individuals resort to deliberate self-injury to cope with their acute feelings. In fact, self-harm is reported to be one of the main causes of death at a young age (Rane and Nadkarni 2014). Since 2005, the percentage of teenagers and young adults who experience major psychiatric distress, including suicidal ideation and suicide attempts, has climbed by 71% (Twenge et al. 2019). In India, around 85% of the total lives lost due to self-harm between 2015 and 2019 belonged to the age group 15 to 49 (Rane and Nadkarni 2014).

Besides identification, mental health support systems face another challenge – people suffering from depression find it hard to express their feelings because they are ashamed and have a constant fear of being judged (Wasserman et al. 2012). Social media platforms like Facebook, Instagram, X (Twitter) provide space for users to connect with others, build communities, share personal life events, spread awareness, and overcome fear and stigma (Ernala et al. 2018). Almost 60% of the Indian population are active on social media (Basuroy 2022a). Apart from entertainment, social media has evolved as a significant resource of information on mental health. Researchers have leveraged social media platforms for analysis and detection of self-harm content (Chancellor, Baumer, and De Choudhury 2019; Aldhyani et al. 2022; Ophir et al. 2020). Social media like X allows users to be pseudo-anonymous (disguised under imaginary profile names), allowing people to reach out to others for help without revealing their actual identity (Coppersmith et al. 2018; Robinson et al. 2015; De Choudhury et al. 2021). This also helps to mitigate the bias present in surveys and self-reported data (Yuan et al. 2023).

While there have been multiple studies on mental health issues observed from the prism of social media (Aldhyani et al. 2022; Ophir et al. 2020), we believe that such studies need to be undertaken in a particular geopolitical and cultural context and in this work, we focus our attention on India. Every nation possesses its own set of challenges, and India is no exception. With a growing economy and the dubious distinction of being the most populous nation with 1.43 billion people, India has been witnessing numerous highs and lows over the last few years (Mozumder 2023; Chatterjee 2023), ranging from the crises of unavailability of hospital beds and medical oxygen during the height of COVID-19 pandemic in May 2021 (Ghoshal 2021), the plight of the migrant workers during nationwide lockdown (Varma 2019) highlighting societal inequalities, to farmers' protest against three agricultural laws.[1] Moreover, often *Dalits* and religious minorities[2] have faced the problem of identity, security, social discrimination, communal tensions and riots (Insights 2022; LotusArise 2022; Nations 2021). Such socio-political activities have great implications on person's men-

[1]The farmers' protest took a toll on the agitating farmers' mental health leading to three suicides (BBC 2021).

[2]In December 2019, the Indian Parliament passed Citizenship Amendment Act (CAA) to provide a route to citizenship to members of six religious communities from Pakistan, Bangladesh and Afghanistan but not for Muslims, which led to nationwide protests spearheaded by Indian Muslims (BBC 2019).

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Issues	Post
Covid Oxygen	My friends father was on the way to hospital and said pls see that I come back home, I dont want to die gasping on the want of oxygen. And on day 3 we get a call that he is no more. These stories will keep on going in our head with pain.
Labour Rights	Here in India - Migrant workers have died and are dying. Poor are dying. Children are suffering. Speak up India. Speak up. #SpeakUpIndia #DeathsofMigrantWorkers #MigrantLaboursDying #SocialRevolution #WorkersUnite #EndHunger #GiveJobs #EqualPay #ConstitutionalRights
Farmer Protest	How do we account for the harms caused to so many farmers and their families during the period preceding the withdrawal? How culpable is the union government and what should be the appropriate reparation? #FarmLawsRepealed #FarmersProtest"
Minority	#Minority communities could not avail govt benefits due to lack of documentation and fear of discrimination. #Muslims fear approaching state institutions for any kind of redressal for the fear of discrimination. No families received scheme benefits in 4%-34% of 53 Muslim areas.
Dalits	Dalit student other than me from our batch was attacked with so many questions in school. And we both had only each other to share these everyday 'casteism' we had to face from school.
CAA	SOS - I am currently being detained at Nagpada Police Station. The police is targeting me now for exercising my fundamental right to peacefully protest against CAA/ NRC. @MumbaiPolice I have not violated any laws. I will take this further to the courts if need be.

Table 1: Tweets highlighting major challenges faced by Indians over past few years (author info omitted for privacy).

tal health. Table 1 shows some example tweets about the major issues that have been surfacing in India over past years.

Earlier works on mental health of Indian social media users (Di Cara et al. 2023; Roy et al. 2021; Lathabhavan 2020; Thippaiah, Nanjappa, and Math 2019; Kumar and Nayar 2021; Barkur, Vibha, and Kamath 2020; Khasawneh et al. 2020) have looked at a binary categorisation: one either has mental health issues or not. However, in reality, people gradually transition from the initial stage of anxiety to depression, leading to extreme steps like suicide (Ageitos et al. 2021). Identification of this transition in the early phase can help in preventing cases of self-harm. Furthermore, changing social, political and personal elements can have an additive effect on a person's mental discomfort. Depression and anxiety level keeps on changing with external factors and thus there is a need to comprehend the change in users behaviour over time. Towards this, in this work, we undertake a longitudinal analysis of Indian social media users, categorising them into three categories based on the posting frequency of afflicting content – low, moderate and high frequency users.

Collecting extensive longitudinal data from X between 2017 and 2021 for all three user categories, we try to answer the following research questions in this paper: How engagement and activity levels vary across user categories? How users across different categories describe themselves? How tweeting preferences, temporal characteristics, and cognitive traits evolve over time for various user categories?

According to our analysis, users who post self-harm content more frequently follow many individuals but have fewer people following them. As we transition from high to low frequency users, the trend seems to reverse. The largest number of followers and lowest number of following are among low frequency users. Most high frequency users connect their identity with their hobbies and passions, while a larger proportion of low and moderate frequency users associate themselves with personal attributes and are vocal about their concerns. We also examine the reactions of other users to tweets posted by self-harm users. In contrast to tweets from low and moderate frequency users, tweets by high frequency users receive more comments. Additionally, self-harm content posted by high frequency users receive more attention as compared to non self-harm content posted by the same user set. High frequency users use mentions to spread the word about their issues, and they are often seen addressing similar topics in their original tweets, retweets, and comments.

Overall, our work helps in uncovering different categories of users undergoing mental health issues and how their social media usage evolves with time. We hope that this work will lead to follow-up efforts to identify early onset of mental health issues and enable support systems to act accordingly.

Related Work

Social media has become a crucial tool for analysing mental health as it is easily accessible (Kosinski et al. 2015), overcoming a number of difficulties in cutting-edge clinical mental health evaluation techniques that depend on subjectivity and retrospective recall bias (Lazer et al. 2009). Considerable work is done in identifying and predicting depression among social media users (Coppersmith et al. 2018; De Choudhury, Counts, and Horvitz 2013; De Choudhury et al. 2021); implication of anonymous nature of platform on self-disclosure (Andalibi et al. 2016; Ernala et al. 2017); comparing the online and offline mental health behaviours (Saha et al. 2017); comprehending social support measures to encourage positive mental health (Andalibi, Ozturk, and Forte 2017; De Choudhury and Kiciman 2017); identifying the norms and practices of community (Chancellor et al. 2016); and how social platforms can be leveraged to provide support (Inkster et al. 2016). Previous studies (Eichstaedt et al. 2018; Owen et al. 2023; Schemer et al. 2020; A. et al. 2022) have conducted longitudinal analysis to predict depression using social media data. However, our research

hang myself	hate myself	am lonely
self injury	take my life	want death
self harm	end my life	kill myself
want to die	been suicidal	cut myself
suicidal thoughts	commit suicide	have no one

Table 2: Keyphrases used for collecting self-harm posts.

specifically targets Indian users, offering a more comprehensive examination of posting behaviors and user traits. To the best of our knowledge, this is the first work that performs longitudinal analysis of self-harm users in India based on the users posting frequency.

Dataset Gathered

Our dataset curation went through the following stages: (i) collect tweets with several keyphrases, (ii) manually annotate them to identify true instance of self-harm and get rid of false positives, (iii) collect timeline of self-harm users, (iv) gather a collection of random tweets, and (v) build and apply a binary classifier to label timeline posts as self-harm or non self-harm. Figure 1 shows various stages of data collection.

Collecting Self-Harm Tweets

We collected all public tweets with geolocation India between 2017 to 2022 which included any of the keyphrases mentioned in Table 2, using the (now defunct) Twitter API with academic research access. These key phrases were curated using the phrases mentioned in prior literature (Coppersmith, Dredze, and Harman 2014; Coppersmith, Harman, and Dredze 2014). Through this data collection process, we obtained 3,759 posts from 3,026 unique users. However, the mere presence of these keyphrases does not necessarily indicate intention of self-harm. Table 3 shows example instances where the presence of a keyphrase can denote both presence and absence of self-harm intent. While there are tweets depicting real cases of anxiety, depression and self-harm; there are also tweets on disappointment with government rules and policies, political and economic situation, death of a famous celebrity, criminal offence or suicide in neighbourhood, suicide awareness and prevention, dialogue from a movie or drama, complaints about customer services etc.

To separate the true instances from false positives, we manually annotated 3,759 posts as valid (true case) or invalid (false case). We employed three annotators, and disagreements were resolved by a majority vote.[3] We find 1,147 posts from 939 users depicting actual instances of mental affliction. Since these 939 users have posted at least one mentally alarming post between 2017 and 2022, we refer to them as the *Focus Group*. We concentrate on this focus group for further analyses.

Compiling X Timelines of the Focus Group

Longitudinal studies have proven to be useful in explaining how people behave in various situations. We analyse the focus group over time from 2017 to 2022. Towards this, we attempted to collect the X timelines of all 939 users from the focus group. We found 49 accounts to be deleted or suspended, and gathered the timeline data for the remaining 890 users. The Twitter API generates an error message 'user not found' or 'user has been suspended' for unavailable accounts, we utilise this information to monitor deleted and suspended accounts. Overall, we obtained 1,385,535 tweets posted by the 890 users of the focus group.

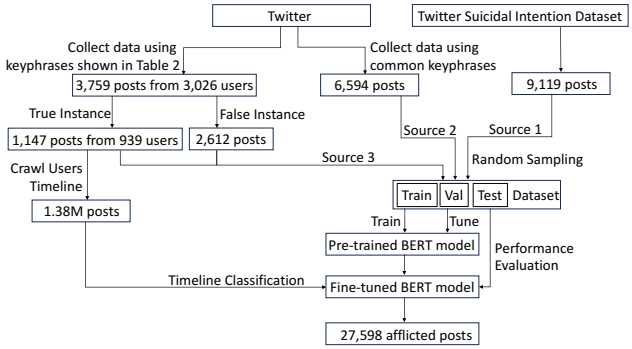


Figure 1: Steps followed for collecting data from users timeline. Transformer based BERT model is fine-tuned using data from three different sources. Model reported F1-score of 0.91 on test dataset. Fine-tuned model classified 27,598 posts as mentally alarming.

Classifier for Automated Labelling

For further analyses, we need to identify how many among these 1.38M tweets are related to self-harm. Since manual annotation of 1.38M tweets is challenging, we use the state-of-the-art transformer based BERT model (Devlin, Chang, and Lee 2019) for the classification task. Previous research has indicated BERT's efficacy in binary classification tasks in the realm of mental health (Owen et al. 2023). For our task, we fine-tune the BERT base model to classify a post as self-harm or non self-harm.[4]

Data for fine-tuning. The data for fine-tuning and testing is collated from three different sources: (i) *Twitter suicidal intention dataset* [5] with 9,119 tweets, amongst which 3,998 posts are labelled as suicidal intention and 5,121 are labelled as non suicidal intention; (ii) we additionally collect 6,594 tweets from X using keywords like 'shopping', 'travel' and 'happy'. Rationale behind this is that timeline posts can deal with routine activities and may not depict self-harm content always, hence such data is necessary for training the classifier;[6] and (iii) the initial manually annotated dataset with 1,147 self-harm tweets and 2,612 non self-harm tweets.

[3]The annotators are aware of the Indian context and have a good understanding of self-harm and non self-harm text.

[4]The classifier code is available at https://github.com/garimachhikara128/MentalHealth/

[5]The dataset is available at https://github.com/laxmimerit/twitter-suicidal-intention-dataset/

[6]We assume that since these tweets are obtained through keywords denoting happy state of mind, they are non self-harm posts.

Self-Harm Intent Present
I've decided to by myself take psychiatric help to give myself one last chance. If I see no improvement then I may just **end my life**.
My life at the current stage is making me to want to **end my life**. I'm slowly losing all purposes that have kept me alive to this day. I don't know how much longer before I can no longer take it and finally **end my life**.
Today's very very bad day, Anniversary of my parents, I'm the very unlucky daughter, I **hate myself** and my worst fate.
Yesterday and for many months before that, I've had intense stages of **self harm**. And these happen late, usually around and after 12 AM. And there's no suicide prevention helpline available. I called up one yesterday because I had self harmed and they asked me to try Savasana ???
Self-Harm Intent Absent
It would have **been suicidal** for congress to let priyanka fight from Varanasi. She is just entering political arena and would have surely lost big from Varanasi. A safe seat like Wayanad is required for her..!
After doing all household chores, I am loving to spend my time on Instagram, YouTube and WhatsApp more these days. I am not much on Facebook because I want to **cut myself** from all the food pics.
It's end of the crossroads now, wanna **end my life** long relationship with @Airtel_Presence, wanna port my SIM into @JioCare, Please let me know the process!
Today at 12:25AM, I have spent more than 45 mins in total waiting and searching for a ride at @Uber_India. I almost **hate myself** for being in situations like these where I need to rely on your service despite of the certainty of your services.

Table 3: Example instances showcasing presence and absence of self-harm intent.

Data	**Original**			**Sampling**			**Train + Validation**			**Test**		
	Class 0	Class 1	Total	Class 0	Class 1	Total	Class 0	Class 1	Total	Class 0	Class 1	Total
D1	5121	3998	9119	2000	3998	5998	1600	3198	4798	400	800	1200
D2	6594	0	6594	2000	0	2000	1600	0	1600	400	0	400
D3	2612	1147	3759	1140	1142	2282	912	913	1825	228	229	457
							4112	4111	**8223**	1028	1029	**2057**

Table 4: Table showcasing equal distribution of training, validation and test data from each data source. *Original* indicates the actual data collected through existing repository and through X API. Random *Sampling* is done to guarantee equal proportionality from both classes. 80% of the sampled data is used for training and rest 20% is used for testing. Amongst training data, 80% is used for fine-tuning the model and other 20% is used for validation. Test data is employed to evaluate the model's performance.

Then, we randomly sample the dataset to ensure an equal number of tweets from both classes, resulting in 8,223 tweets in the training+validation dataset and 2,057 tweets in the test dataset. Table 4 shows the detailed breakup of posts used for training, validation and testing. We observe that the final fine-tuned BERT model exhibits good performance on the test dataset with an accuracy of 91%. After applying this classifier to 1.38M posts collected from different users timelines, 27,598 tweets are classified as self-harm and the remaining tweets are classified as non self-harm.

Categorising Users in the Focus Group

Based on the frequency of self-harm posts over the years, we categorise the users from the focus group into three categories – **low frequency users**, **moderate frequency users**, and **high frequency users**. Figure 2 shows the frequency of self-harm posts between 2017 and 2022 posted by all users. We define low frequency as those who posted less than 50 self-harm posts, moderate frequency who posted more than 50 but less than 150 tweets, and high frequency who posted more than 150 tweets. Details about the user distribution across these categories are shown in Table 5. We observe that 1%, 3% and 8% of tweets posted by low, moderate and high frequency users respectively reveal the presence of mental discomfort. This suggests that when compared to low and moderate frequency users, high frequency users not only post more self-harm content but also a bigger percentage of their posts are distressing.

Characterising Different User Categories

In this section, we examine the long-term trends among different user categories based on social media neighbourhood, bio description, tweeting preferences and temporal variations. While there have been multiple works focusing on mental health concerns, to the best of our knowledge, this is the first work to consider the longitudinal differences amongst Indian users based on the posting frequency.

Social Media Neighbourhood

The number of followers and followees of a user reveal their attempt at socialising as well as the willingness to consume outside information. Earlier works have shown that power law best fits the in-degree distribution (followers), whereas log-normal best fits the out-degree distribution (followees) (Myers et al. 2014; Lerman, Yan, and Wu 2016). Interestingly, in our case (as shown in fig. 3), we observe that

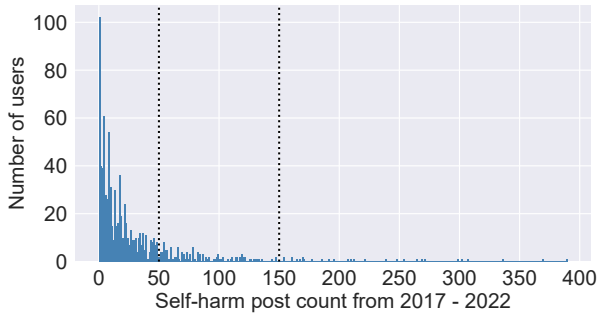


Figure 2: Total number of self-harm posts made by 890 users of the focus group from 2017 - 2022. We select 50 and 150 as threshold for low, moderate and high frquency users. Users with less than 50 self-harm posts are low frequency users, users with self-harm post count between 50 and 150 are moderate frequency users, and users with self-harm post count greater than 150 are high frequency users.

Category	No. of Users	No. of Posts	No. of SH Posts	Fraction of SH Posts
low	739	1023919	10892	0.011
moderate	122	286042	10086	0.035
high	29	75574	6620	0.088

Table 5: Distribution of self-harm (SH) users and posts across various categories.

both follower and followee counts obey log-normal distribution for all three user categories in the focus group. Table 6 shows the value of exponent α for power law and value of mean and standard deviation i.e., μ and σ for log-normal distribution. Moderate frequency users are reported to have the highest mean μ across both the measures of followers and following and it is interesting to observe that the standard deviation σ is the lowest across medium frequency users, indicating that most of the followers and following of moderate frequency users are clustered around the mean, whereas for low and high frequency users the spread is relatively higher as compared to moderate frequency users.

Engagement. To assess levels of engagement across three user categories, we establish three measures based on the volume of users followers, following and tweets (Wang et al. 2017). We characterise the level of engagement of category c in terms of measure m as follows:

$$Engagement(c, m) = \frac{\sum_{u=1}^{n} \#m_u}{n} \quad (1)$$

where $c \in \{Low, Moderate, High\}$ and $m \in \{Followers, Following, Tweets\}$, n are the number of users in category c and $\#m_u$ denotes the count of measure m for user $u \in c$.

Table 7 lists the statistics for social engagement. Low frequency users show the maximum engagement for #followers and high frequency users exhibit the maximum engagement

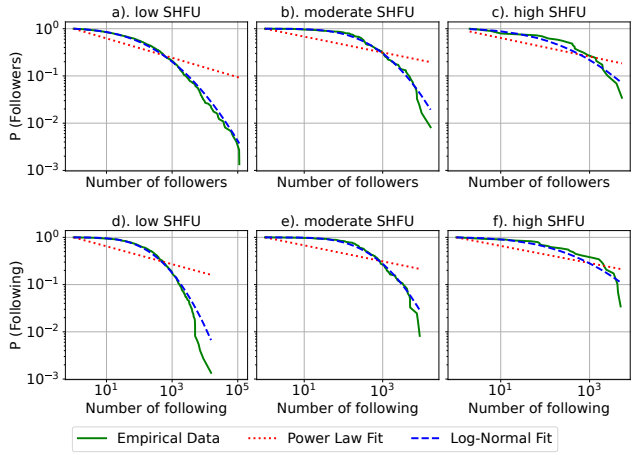


Figure 3: Complementary cumulative distribution function (CCDF) for count of followers and following. Red and blue curves denote fitted power law and lognormal distributions across low, moderate and high self harm frequency users.

Measure	Category	Power Law α	Log-normal μ	Log-normal σ
Followers	low	1.205	4.661	2.589
	moderate	1.168	**5.955**	**1.820**
	high	1.197	4.948	2.456
Following	low	1.189	5.272	1.766
	moderate	1.169	**5.903**	**1.673**
	high	1.181	5.415	2.521

Table 6: Statistics for distribution of followers and following. The parameter α assumes $P(x) \sim x^{-\alpha}$ for degree x (power law). The μ and σ parameters assume $P(x) \sim \frac{1}{x}\exp[\frac{(\ln x-\mu)^2}{2\sigma^2}]$ (log-normal).

for #following and #tweets. One might argue that it was implied for high frequency users to receive the maximum engagement on #tweets since this itself was the categorisation approach, but note that our method of categorisation is exclusively based on the number of **self-harm posts** regardless of the total number of posts. In addition to posting more instances of self-harm, high frequency users have the highest #tweets across all user types. Earlier works have reported the shrinkage of networks as a useful indicator of impending behavioural changes (De Choudhury, Counts, and Horvitz 2013). In contrast to low frequency users, who have a bigger network of #followers, high frequency users have less individuals following them. This suggests that users who publish more distressing content are less likely to be followed by other users. High frequency users have highest #following, which indicate they are more intrigued about others as compared to low and moderate frequency users. Kolmogorov-Smirnov (KS) test (Lilliefors 1967) revealed the results to be statistically significant with $p < 0.05$.

Activity. We investigate which user category is most active on Twitter. We quantify activity as the average number

Measure	Low	Moderate	High
Engagement			
#Followers	**1749.51**	1372.76	817.07
#Following	643.84	1035.57	**1203.97**
#Tweets	1385.55	2344.61	**2606.00**
Activity			
#Followers/day	0.65	**0.72**	0.37
#Following/day	0.61	0.65	**0.94**
#Tweets/day	1.05	3.06	**11.1**

Table 7: Social Media Neighbourhood: Statistics for Engagement and Activity across various user categories.

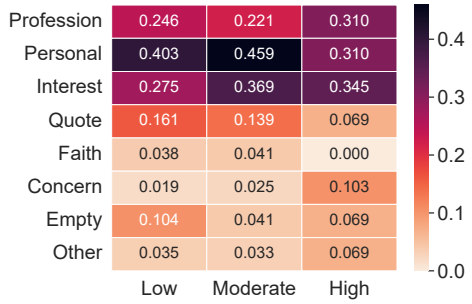


Figure 4: Normalised score for labels across three different user categories.

of followers, following and tweets per day.

$$Activity(c, m) = \frac{\sum_{u=1}^{n} \left(\frac{\#m_u}{d_u} \right)}{n} \quad (2)$$

where d_u is the number of days between – user u joining X and his last post. Higher activity levels indicate higher expressiveness (Ernala et al. 2017; Saha, Weber, and De Choudhury 2018; De Choudhury et al. 2021; Yuan et al. 2023). Moderate frequency users have active #Followers/-day whereas #Followers are not the highest for them, this indicate users with moderate frequency are engaged for a shorter amount of time on Twitter.

Bio Description

Users embark on significant communicative and identity disclose properties when they create their unique accounts, e.g., by filling up the bio sections, as opposed to being confined solely to the content that gets posted thereafter (Greene and Brownstone 2023; Chakraborty et al. 2017). We analyse bio description of the 890 users from the focus group and label them as 'profession', 'personal', 'interest', 'quote', 'faith', 'concern', 'empty', and 'other'.

Profession: where a user describes something about their profession, e.g., *"Associate Director"*, *"Practicing Gynaecologist"*, *"Assistant Professor(Pharmacology)"*, *"Fitness Coach"*.

Personal: information regarding age, gender, birthday, education, location, religion, qualities and relations, e.g., *"she/her 24"*, *"Astrophysics Scholar"*, *"Believe in Secularism — Peace — Love"*, *"A proud Resident of Bihar"*, *"Proud Father"*.

Interest: where users highlight interests or hobbies, e.g., *"BTS : Justin"*, *"Interest in Photography & Documentary Making"*, *"a singer and baker by hobby"*.

Quote: users use well known quotes or dialogue from movies or pen down their thoughts in bio description, e.g., *"Emptiness is better than something temporary"*, *"When life gives you lemons, make lemonade of it"*.

Faith: user mention about religion and God, e.g.,*"God Protect Me"*, *"Jai Shree Ram"*, *"Inservice to Lord Krishna"*, *"Burn the evil spirits and demons now and forever. Amen omshanthi Allah"*.

Concern: user raises concern regarding personal or societal issues, e.g., *"#cancel 12th board exam"*, *"Alone boy. Need adoption or family"*, *"Dream IIT but no money and guidance. Please help me for higher education."*, *"My life Dedicated towards Saving & Serving Innocent Speechless Souls (Animals)..In real,trying to be humane"*.

Empty: user prefers to leave their bio description unfilled.

Other: bio description is not significant enough to convey meaning, e.g., *'#'*, *'eh.'*, *'...'*, *'anyway,,,'*.

People use bio to express their *self-identification*, and individuals can opt to do so from a variety of perspectives, hence a bio description can be tied to several labels. Figure 4 shows the normalised count of labels across different user categories. 40.3% of low frequency and 45.9% of high frequency users mention about their *personal* lives in their bio descriptions, followed by *profession* and *interest*. Majority of the low and moderate frequency users discuss their personal traits on social media, on the contrary only 31% of high frequency users mention personal elements, which suggests fewer high frequency users prefer to divulge their personal identities on social media. The biggest percentage of high frequency users (34.5%) mention their amateur interest, suggesting that they possess or at least identify themselves with some leisure pursuits. An equal percentage (31%) of high frequency users mention about *profession* and *personal* elements, among high frequency users *identification* is nearly equally distributed across *profession*, *personal* and *interest*. Faith is mentioned in 3.8% and 4.1% of user bios from low and moderate frequency users respectively, on the contrary none of high frequency users referred to or exhibited trust in God in their bio descriptions. Low and moderate frequency users have only 1.9% and 2.5% bio descriptions tied to *concern*, whereas, among high frequency users 10.3% of tweets address critical issues, suggesting high frequency users are relatively more vocal about venting concerns.

Tweeting Preference

We analyse how users from the focus group engage with one another – interaction with posts and publishing of posts (Wang et al. 2017).

Post Interaction. We consider three measures for interaction with post - replies, likes and mentions. Figure 5 shows average number of replies, likes and mentions received on ***all set*** and ***self-harm set*** for different user categories.

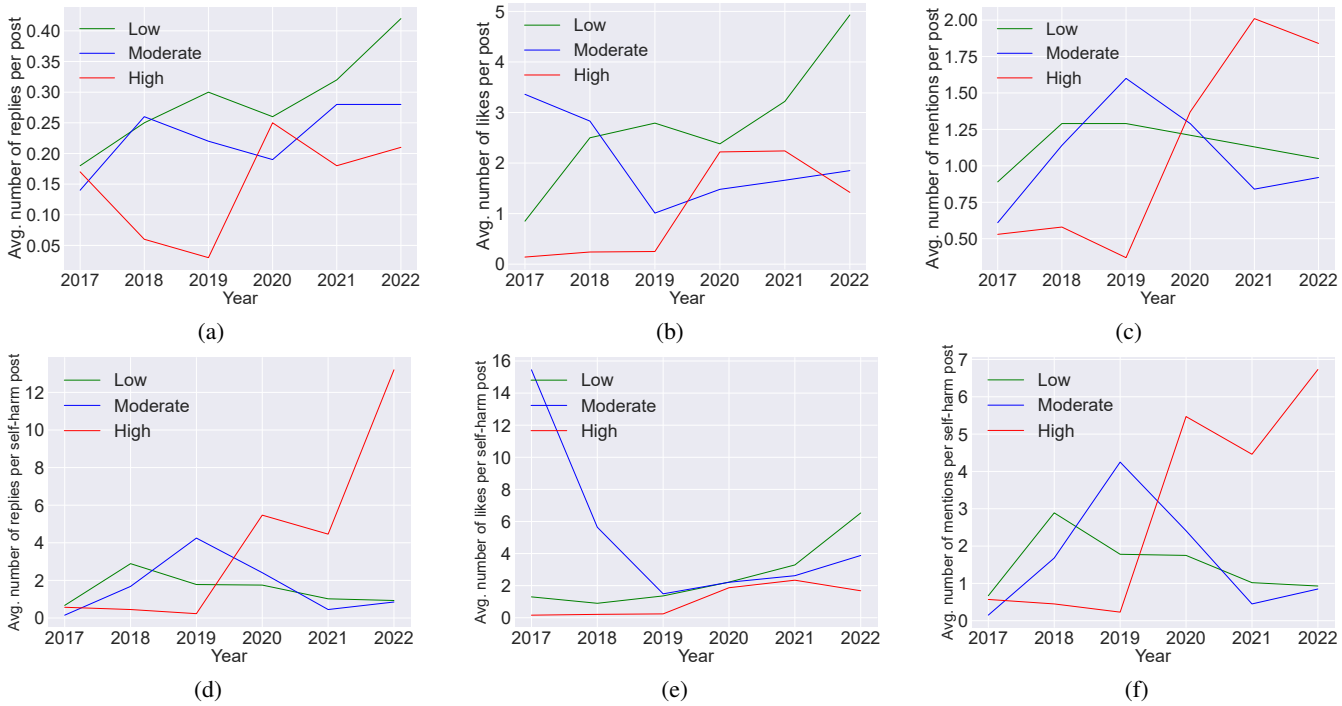
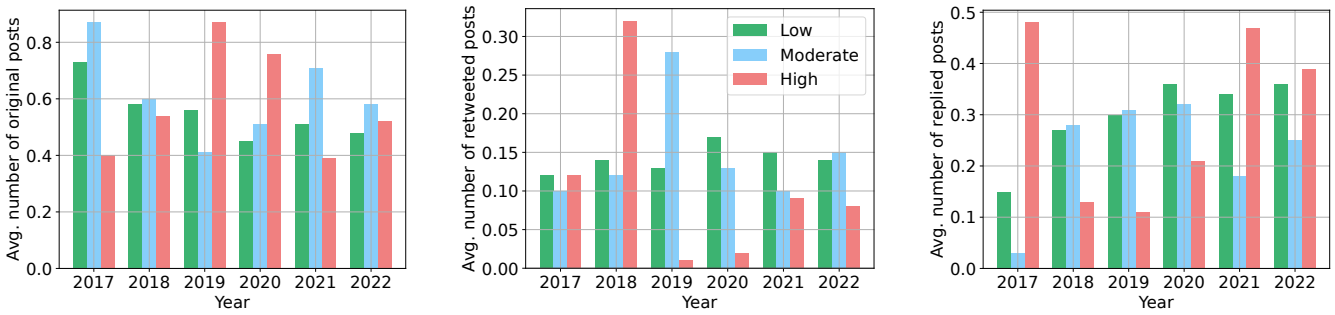


Figure 5: Post Interaction from 2017–2022 for various user categories. Average number of (a) replies (b) likes (c) mentions received on all set. Average number of (d) replies (e) likes (f) mentions on self-harm set.

We first compare the measures across all set and self-harm set. All set comprises of the entirety of timeline posts made by users of the focus group, self-harm set refers to mentally afflicting posts from the focus group's timeline. i) Replies for all set vary between [0,0.4] (fig. 5a), whereas for self-harm set replies range between [0,12] (fig. 5d). Self-harm set exert higher interaction – the average number of replies are greater in comparison to all set, people on social media exhibit higher interaction with self-harm posts indicating higher social assistance, some earlier works have shown replies can act as a proxy for social support (De Choudhury and Kiciman 2017). ii) Number of likes for all set varies between [0,5] (fig. 5b), whereas the spread for self-harm set is [0,16] (fig. 5e), however if we consider year 2018 and later, the number of likes drop to range of [0,7], which is sufficiently close to the likes received on all set. If we investigate recent years, we can deduce that all set and self-harm set receive a nearly identical pattern of interaction with regard to likes. iii) Figure 5c and 5f illustrates the mention count on all set to range between [0,2] and self-harm set to range in [0,7] respectively. Self-harm set have a higher count of mentions as compared to other posts, which indicates the intent to disseminate the self-harm post to a broader audience.

Secondly, we compare user categories across various measures. i) Replies. On all set, low frequency users receive highest replies and high frequency users seek minimum replies (fig. 5a), whereas the trend seems to be reversed when we consider the self-harm set – high frequency users gather maximum and low frequency users receive minimum replies (fig. 5d). This reveals that people on social media converse more with low frequency users when they post non self-harm content and people engage more with high frequency users when they post self-harm content. ii) Likes. Over recent years, from 2019 – 2022, the pattern of likes across user categories has been similar for all set and self-harm set (fig. 5b and 5e), low frequency users receive maximum number of likes whereas high frequency users obtain minimum number of likes. This indicates social media users do not prefer interacting with high frequency users through likes. iii) Mentions. Usage of mention implies the want to let the other party know about the problem or anticipate a solution. Mention for high frequency users increased drastically since 2019, high frequency users avail the maximum usage of mentions across all set and self-harm set (fig. 5c and 5f) indicating they want their posts to reach to the maximum audience, and in haste, the users tag all potential helpers. While we observe an increase in mentions for high frequency users since 2019, there is a dip in mentions for low and moderate frequency users across all set and self-harm set. KS test indicates statistical significance for results shown in Figure 5 with $p < 0.001$.

The takeaway is high frequency users garner maximum replies for self-harm posts, whereas for general posts low frequency users attract the most replies. For year 2018 – 2022, statistics for likes and mentions follow similar patterns for various user categories across all set and self-harm set. Likes received by a user on a normal post and on a self-harm post are alike, and likes are not the preferred way of interacting with high frequency users. Self-harm posts comprise higher mentions as compared to general posts.



(a) (b) (c)

Figure 6: Fraction of (a) original (b) retweeted and (c) replied posts in all set from 2017 - 2022 for various user categories.

Post Type	User Category	Distinct Frequent Keywords
original	low	cope, feelings, depressed, stressful, crying, unhappiness, emotionless, mistreatment, crying, attachment, loneliness, distress, longing, grief, stress
	moderate	existence, imprisons, calm, silenced, deprivation, isolate, relativesam, memories, suicide, psychopath, awareness, hide, disappear, paranoia, invisibility
	high	admission, prevention, exams, intervene, discrimination, depression, anxiety, stressed, sufferings, hospitalisation, cbse, exams, pandemic, students, covid
retweet	low	dharmapuri, prayer, spiritual, coping, depressed, treatment, depression, patanjali, psychotherapy, pranayama, ayurveda, meditating, buddha, yoga, depressant
	moderate	meditation, addictive, smoking, overcome, remorse, suicides, mindfulness, therapy, psychiatric, depressed, addict, debilitating, suicidal, craving, medications
	high	study, stress, grades, delaying, appeal, student, universities, silence, education, tuition, protest, campus, prevention, admission, exam
reply	low	abused, oppression, slander, threatening, crimes, murders, vaccination, grievance, assaulted, violence, hypocrisy, pervert, victim, rapists, molestation
	moderate	husbands, disorders, loneliness, marriage, dysfunction, introvert, societal, depression, stigma, divorce, therapy, coping, patriarchal, matrimonial, feminists
	high	castes, student, cbse, delhi, panchayat, conferred, bharat, exam, declared, india, xii, assessment, mandir, examination

Table 8: Distinct high frequency keywords in original posts, retweets and replies.

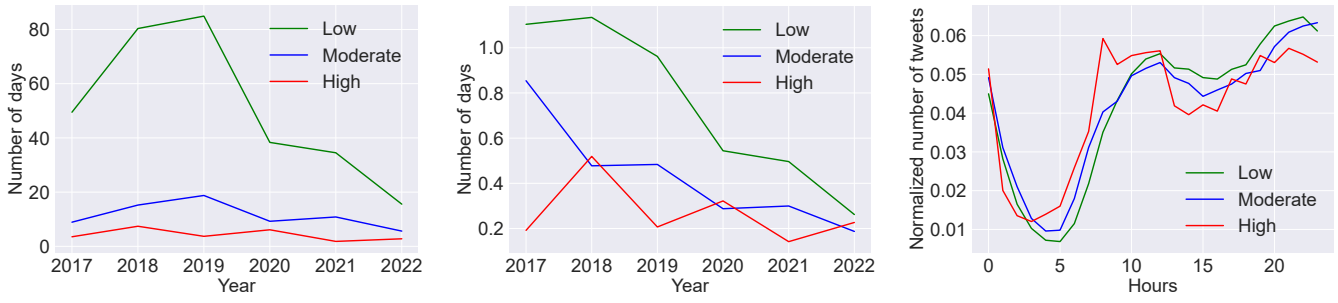
Publishing Posts. Users can publish posts in three ways: original post, retweet, or reply. We analyse the timeline posts of three user categories. Original post initiates a new discussion or thread, retweeted post is the retweet of another post, and replied post is the comment provided on another post.

In 2021 and 2022, all three user categories posted more original content followed by replied and retweeted posts (fig. 6). From 2017 - 2022, retweeted posts are observed to be minimal across all user categories, which indicate retweet is not extensively used by users.

We find the most frequently used words in original posts, retweets and replies across different user categories (table 8). We leverage the use of KeyBERT for keyword extraction, KeyBERT utilises BERT embeddings to generate keyphrases and keywords that are the closest to the given document (Grootendorst 2020). New conversation initiated by a low frequency user mostly talks about depression, stress, anxiety, and grief whereas a moderate frequency user posts about invisibility, isolation, and question their existence. In retweets, users suggest methods for coping with various mental issues through meditation, pranayama which is a breath regulation exercise, yoga, prayer, and medication. In replies, users are more expressive about their personal lives and share their perspectives about society, family, and personal matters like violence, divorce, and stigma. High frequency users, across all the posts, often discuss about exams, studies, students, results, and grades; this section of self-harm users seems to be mostly students and are trying to raise their concerns via new threads, retweets and replies. There is variation in the posts made by low and moderate frequency users across new posts, retweets and replies. Whereas high frequency users seem to be constantly discussing about the same issue across all the posts.

Temporal Variation

We study the relation between user categories and variation with time, we utilise the entire timeline data of 1.38M posts for temporal analysis. Low frequency users report highest average time gap between any two successive self-harm post, followed by moderate and high frequency users (fig. 7a). One might argue, this observation was implied as low frequency users post the least amount of self-harm posts, to



(a) (b) (c)

Figure 7: (a) Average time gap between two successive self-harm post (b) average time gap between any two successive post (c) time of the day when a post is created.

answer this, let us consider an example, a user posts three self-harm posts on day 1, day 4 and day 11 respectively, average time gap between two successive posts is 5 days. On the contrary, if the user posts on day 1, day 7 and day 15, the average time gap will be 7 days. In both cases, user posted three tweets but the average time gap differs, our method of categorisation is solely dependent on the number of self-harm posts. High frequency users tend to post self-harm content more frequently.

Similar statistics are observed for the average time gap between any two successive posts, low frequency users are reported to have the highest time gap (fig. 7b), which however has tremendously decreased over the past years. Low frequency users are less involved towards self-harm and general posts. On the other hand, high frequency users exhibit elevated levels of activity for both self-harm and general posts. If we look at the broader pattern, since 2017, the average time gap on self-harm posts has decreased for all three user categories, indicating people are getting more engaged and responsive to posts. KS test revealed results of fig. 7a and 7b to be statistically significant with $p < 0.001$.

We analyse the time of the day when users post. People with mental health issues may experience insomnia, they might act differently from regular users in terms of their online activities (Lustberg and Reynolds 2000). High frequency users are more engaged as compared to low and moderate frequency users from 3 am to 9 am and are least active from 11 am to 7 pm (fig. 7). Sleep cycle is particularly worse for high frequency users, who appear to be most active in early mornings and less active during the day. Prior work has suggested that disrupted sleep is one of the factors that is closely associated with depressive illnesses (Abdel-Khalek 2004). High frequency users are most prolific during odd hours which raises concerns about their mental health.

Cognitive Attributes

We study cognitive measures such as readability, complexity and repeatability for various user categories in all set.

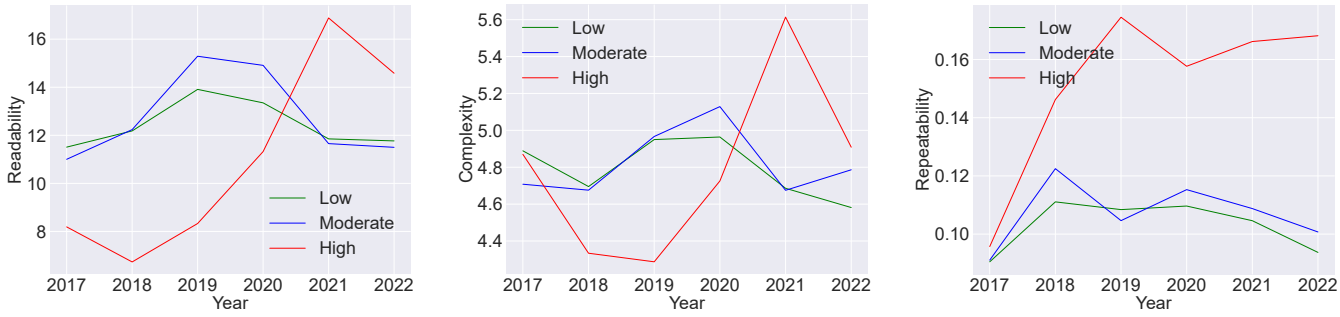
Readability. Readability is a measure to gauge the ease with which readers may understand a certain text (McCallum and Peterson 1982). Readability is an important indicator of people's cognitive behaviour, and earlier research has utilised this metric to comprehend the patterns of conversation in social networks (Ernala et al. 2017; Saha, Weber, and De Choudhury 2018; Saha and Sharma 2020). We employ Coleman-Liau Index (CLI) to calculate readability. CLI is calculated as, $CLI = 0.0588 * L - 0.296 * S - 15.8$, where L is the average number of letters per 100 words and S is the average number of sentences per 100 words (Pitler and Nenkova 2008).

We analyse the readability index in all set for different user categories (fig. 8a). CLI for low and medium frequency users are observed to be greater than 11 throughout the timeline of six years, depicting that the data is fairly difficult to read. For high frequency users CLI was lower than 8 until year 2019, indicating the text is ideal for average readers, but since then CLI has increased and in recent years is found to be higher than low and moderate frequency users. This indicates, as opposed to earlier years, high frequency users are now expressing their thoughts in fairly complex language.

Complexity. Complexity is the average length of words per sentence (Ernala et al. 2017; Saha and Sharma 2020). Since 2019, there has been an increase in the complexity for high frequency users with a decline in complexity for low and medium frequency users (fig. 8b).

Our analysis reveals that, since 2019, for low and medium frequency users there is a decrease observed in readability and complexity, and there is an increase for high frequency users. Earlier works (Ernala et al. 2017) have shown psychosocial health to have a positive correlation with readability and complexity. Note that, the data set used by (Ernala et al. 2017) is from year 2012 to 2016, which is disjoint with our data set which is from 2017 to 2022. If we base our analysis on the arguments made by (Ernala et al. 2017), from year 2017 to 2019, high frequency users show low scores for readability and complexity, indicating lower well-being. With many new Indian people joining X over the past few years (Basuroy 2022b), posting behaviour has changed substantially, high frequency users form more complex sentences as compared to low and moderate frequency users.

Repeatability. Repeatability is the normalised count of non-unique words, and correlates negatively with psychosocial health (Ernala et al. 2017; Saha, Weber, and De Choud-



(a) (b) (c)

Figure 8: (a) Readability (b) Complexity (c) Repeatability scores for all set across different user categories.

hury 2018). For high frequency users, repeatability is observed to be higher as compared to low and moderate frequency users (fig. 8), and higher repeatability is indicative of low well-being.

Takeaway. Since 2020, high frequency users from India have shown an increase in readability, complexity and repeatability, indicating they form more complex sentences and repeat words as compared to low and moderate users.

Concluding Discussion

We gathered timeline information from 1.38 million posts for 890 self-harm users in India. We opted for an automated transformer-based BERT model to classify 1.38 million tweets as self-harm and non-self-harm. On the basis of the number of self-harm tweets, we suggest a system to categorise self-harm users into three categories: low, moderate, and high. We conducted a number of analysis to learn more about the mental health of users in different categories. According to our findings, high frequency users are the most active; they follow many individuals and, in contrast, obtain fewer followers. High frequency users have highest comment counts, and likes are not the preferred method of communication with them. High frequency users had the worst sleep patterns of all and tend to write long, complicated phrases with repeated words. High frequency users show evidence of poor mental health, this category of users require quick assistance. The results of this study can provide assistance in early healthcare intervention in cases of self-harm.

Implications

Rich literature on mental health has considered two categories of users - one who face anxiety and depression and the other who does not. It is important not to place the users in either of the two axes – positive or negative. People who post mentally afflicting content can have different levels of affliction, motivated via this, we proposed a method to classify self-harm users in different categories based on their posting frequencies. Analysis reveals the psychosocial well-being of users from different categories, and the longitudinal analysis informs about the changing patterns across various measures.

Our work has practical implications for preventing self-harm. We currently assign each user a fixed category, but this work can be modified to allow for dynamic categorisation, where users who appear to change their categories from low to moderate or from moderate to high require immediate attention and users who change their categories from moderate to low or high to moderate can be seen benefiting from social media.

Limitations and Future Work

We acknowledge the limitations of our work, some of which suggest interesting areas for additional investigation. Our analysis is probably affected by selection bias, we collect publicly available data from X, given the stigma associated with mental illness (Corrigan 2004), only a small fraction of people prefer to post their opinions openly on social media. Our dataset encompasses the timeframe from 2017 to 2022, which coincided with the COVID-19 pandemic. This global event profoundly affected individuals' lives, with many experiencing personal losses. As a result, the user behavior depicted in our findings may be influenced by this unique period. It is imperative to recognise that conducting a similar study during a different timeframe could yield varying behavioral outcomes.

While India boasts a rich linguistic diversity, our analysis is confined to English tweets. It's noteworthy that many Indians utilise *code-mixing* (such as, Hinglish) for their social media interactions, suggesting a promising avenue for exploring tweet analysis across diverse Indian languages. Furthermore, our analysis overlooks user demographics. Individuals of varying genders, ages, and locations encounter distinct challenges. Delving into user characteristics at a more granular level, rather than merely at the national level, presents an intriguing opportunity for future research.

Ethics Statement

We utilise publicly available data from Twitter and we commit to safeguard the privacy of individuals. In this paper, we prioritise anonymity by removing all personal identity information and rephrasing the Twitter posts included in this document to prevent identification. Nevertheless, we recognise the potential negative impact of this work. Reporting of user characteristics of different categories may lead to

high anxiety levels among other users who may experience similar pattern. Moreover, there exists a potential for misuse of the findings presented in this paper, particularly in the development of commercial tools aimed at detecting levels of depression. Such tools could be exploited in various domains, including targeted hiring and insurance premium calculation, thereby amplifying existing societal inequities.

Acknowledgements

We would like to express our sincere gratitude to the anonymous reviewers for their valuable feedback which helped in improving the quality of the paper. We also extend our appreciation to our colleagues Kausik Hira and Vijay Kumar Meena for assistance during the early stages of this work.

References

A., A.; Chaudhuri, R.; Hussain, Z.; and Chatterjee, S. 2022. Social media usage and its impact on users' mental health: a longitudinal study and inputs to policymakers. *International Journal of Law and Management*.

Abdel-Khalek, A. 2004. Can somatic symptoms predict depression? *Social Behavior and Personality: An international journal*.

Ageitos, E. C.; Fabregat, H.; Araujo, L.; and Martínez-Romo, J. 2021. NLP-UNED at eRisk 2021: self-harm early risk detection with TF-IDF and linguistic features. In *CLEF*.

Aldhyani, T. H. H.; Alsubari, S. N.; Alshebami, A. S.; Alkahtani, H.; and Ahmed, Z. A. T. 2022. Detecting and Analyzing Suicidal Ideation on Social Media Using Deep Learning and Machine Learning Models. *International Journal of Environmental Research and Public Health*.

Andalibi, N.; Haimson, O. L.; De Choudhury, M.; and Forte, A. 2016. Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity. In *CHI*.

Andalibi, N.; Ozturk, P.; and Forte, A. 2017. Sensitive Self-Disclosures, Responses, and Social Support on Instagram: The Case of #Depression. In *CSCW*.

Barkur, G.; Vibha; and Kamath, G. B. 2020. Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. *Asian journal of psychiatry*.

Basuroy, T. 2022a. India: Social Network penetration 2025.

Basuroy, T. 2022b. Number of Twitter users in India from 2013 to 2022.

BBC. 2019. Citizenship Amendment Bill: India's new 'anti-Muslim' law explained.

BBC. 2021. Farm laws: India farmers end protest after government accepts demands.

Chakraborty, A.; Sarkar, R.; Mrigen, A.; and Ganguly, N. 2017. Tabloids in the era of social media? understanding the production and consumption of clickbaits in twitter. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW): 1–21.

Chancellor, S.; Baumer, E. P. S.; and De Choudhury, M. 2019. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *CSCW*.

Chancellor, S.; Lin, Z.; Goodman, E. L.; Zerwas, S.; and De Choudhury, M. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *CSCW*.

Chatterjee, R. 2023. India's population passes 1.4 billion — and that's not a bad thing.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*.

Coppersmith, G.; Harman, C.; and Dredze, M. 2014. Measuring Post Traumatic Stress Disorder in Twitter. *ICWSM*.

Coppersmith, G.; Leary, R.; Crutchley, P.; and Fine, A. 2018. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical informatics insights*.

Corrigan, P. 2004. How stigma interferes with mental health care. *The American psychologist*.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Predicting Postpartum Changes in Emotion and Behavior via Social Media. In *CHI*.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2021. Predicting Depression via Social Media. *ICWSM*.

De Choudhury, M.; and Kiciman, E. 2017. The Language of Social Support in Social Media and Its Effect on Suicidal Ideation Risk. *ICWSM*.

Devlin, J.; Chang, M.-W.; and Lee, K. 2019. Google, KT, Language, AI: BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Di Cara, N. H.; Maggio, V.; Davis, O. S. P.; and Haworth, C. M. A. 2023. Methodologies for Monitoring Mental Health on Twitter: Systematic Review. *Journal of medical Internet research*.

Eichstaedt, J. C.; Smith, R. J.; Merchant, R. M.; Ungar, L. H.; Crutchley, P.; Preoţiuc-Pietro, D.; Asch, D. A.; and Schwartz, H. A. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences of the United States of America*.

Ernala, S. K.; Labetoulle, T.; Bane, F.; Birnbaum, M. L.; Rizvi, A. F.; Kane, J. M.; and De Choudhury, M. 2018. Characterizing Audience Engagement and Assessing Its Impact on Social Media Disclosures of Mental Illnesses. *ICWSM*.

Ernala, S. K.; Rizvi, A. F.; Birnbaum, M. L.; Kane, J. M.; and De Choudhury, M. 2017. Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia. *CSCW*.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*.

Ghoshal, D. 2021. Why India is facing an oxygen crisis as COVID cases mount.

Greene, A. K.; and Brownstone, L. M. 2023. "Just a place to keep track of myself": eating disorders, social media, and the quantified self. *Feminist Media Studies*.

Grootendorst, M. 2020. KeyBERT: Minimal keyword extraction with BERT.

Inkster, B.; Stillwell, D.; Kosinski, M.; and Jones, P. 2016. A decade into Facebook: where is psychiatry in the digital age? *Lancet Psychiatry*.

Insights. 2022. Issues related to minorities.

Khasawneh, A.; Chalil Madathil, K.; Dixon, E.; Wiśniewski, P.; Zinzow, H.; and Roth, R. 2020. Examining the Self-Harm and Suicide Contagion Effects of the Blue Whale Challenge on YouTube and Twitter: Qualitative Study. *JMIR Ment Health*.

Kosinski, M.; Matz, S. C.; Gosling, S. D.; Popov, V.; and Stillwell, D. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American psychologist*.

Kumar, A.; and Nayar, K. R. 2021. COVID 19 and its mental health consequences. *Journal of Mental Health*.

Lathabhavan, R. 2020. People and social media platforms for positive mental health- A paradigm shift: A case on COVID-19 impact form India. *Asian journal of psychiatry*.

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabasi, A.-L.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D.; and Van Alstyne, M. 2009. Life in the network: the coming age of computational social science. *Science*.

Lerman, K.; Yan, X.; and Wu, X.-Z. 2016. The "majority illusion" in social networks. *PLoS One*.

Lilliefors, H. W. 1967. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association*.

LotusArise. 2022. Problems of Religious Minorities in India.

Lustberg, L.; and Reynolds, C. F. 2000. Depression and insomnia: questions of cause and effect. *Sleep medicine reviews*.

McCallum, D. R.; and Peterson, J. L. 1982. Computer-Based Readability Indexes. In *Proceedings of the ACM '82 Conference*. ACM.

Mozumder, S. 2023. The World Bank In India.

Myers, S. A.; Sharma, A.; Gupta, P.; and Lin, J. 2014. Information Network or Social Network? The Structure of the Twitter Follow Graph. In *WWW*.

Nations, U. 2021. The Dalit: Born into a life of discrimination and stigma.

Ophir, Y.; Tikochinski, R.; Asterhan, C. S. C.; Sisso, I.; and Reichart, R. 2020. Deep neural networks detect suicide risk from textual facebook posts. *Scientific Reports*.

Owen, D.; Antypas, D.; Hassoulas, A.; Pardiñas, A. F.; Espinosa-Anke, L.; and Collados, J. C. 2023. Enabling early health care intervention by detecting depression in users of web-based forums using language models: Longitudinal analysis and evaluation. *JMIR AI*.

Pitler, E.; and Nenkova, A. 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. In *EMNLP*.

Rane, A.; and Nadkarni, A. 2014. Suicide in India: a systematic review. *Shanghai archives of psychiatry*.

Robinson, J.; Cox, G.; Bailey, E.; Hetrick, S.; Rodrigues, M.; Fisher, S.; and Herrman, H. 2015. Social media and suicide prevention: a systematic review. *Early intervention in psychiatry*.

Roy, A.; Singh, A. K.; Mishra, S.; Chinnadurai, A.; Mitra, A.; and Bakshi, O. 2021. Mental health implications of COVID-19 pandemic and its response in India. *International Journal of Social Psychiatry*.

Saha, K.; Chan, L.; De Barbaro, K.; Abowd, G. D.; and De Choudhury, M. 2017. Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*

Saha, K.; and Sharma, A. 2020. Causal Factors of Effective Psychosocial Outcomes in Online Mental Health Communities. *ICWSM*.

Saha, K.; Weber, I.; and De Choudhury, M. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations after Student Deaths on College Campuses. *ICWSM*.

Schemer, C.; Masur, P. K.; Geiß, S.; Müller, P.; and Schäfer, S. 2020. The Impact of Internet and Social Media Use on Well-Being: A Longitudinal Analysis of Adolescents Across Nine Years. *Journal of Computer-Mediated Communication*.

Thippaiah, S. M.; Nanjappa, M. S.; and Math, S. B. 2019. Suicide in India: A preventable epidemic. *The Indian journal of medical research*.

Twenge, J. M.; Cooper, A. B.; Joiner, T. E.; Duffy, M. E.; and Binau, S. G. 2019. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005-2017. *Journal of abnormal psychology*.

Varma, S. 2019. Govt. Report Reveals Shocking Condition of Workers in India.

Wang, T.; Brede, M.; Ianni, A.; and Mentzakis, E. 2017. Detecting and Characterizing Eating-Disorder Communities on Social Media. In *WSDM*.

Wasserman, C.; Hoven, C. W.; Wasserman, D.; Carli, V.; Sarchiapone, M.; Al-Halabí, S.; Apter, A.; Balazs, J.; Bobes, J.; Cosman, D.; Farkas, L.; Feldman, D.; Fischer, G.; Graber, N.; Haring, C.; Herta, D. C.; Iosue, M.; Kahn, J.-P.; Keeley, H.; Klug, K.; McCarthy, J.; Tubiana-Potiez, A.; Varnik, A.; Varnik, P.; Ziberna, J.; and Poštuvan, V. 2012. Suicide prevention for youth–a mental health awareness program: lessons learned from the Saving and Empowering Young Lives in Europe (SEYLE) intervention study. *BMC Public Health*.

Yuan, Y.; Saha, K.; Keller, B.; Isometsä, E. T.; and Aledavood, T. 2023. Mental Health Coping Stories on Social Media: A Causal-Inference Study of Papageno Effect. In *TheWebConf*.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes
   (b) Have you provided justifications for all theoretical results? Yes
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? Yes
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA, we collected publicly available data from X.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT) | 979-8-3503-7131-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/InCACCT61598.2024.10551077

Leveraging MediaPipe and YOLO Keypoint Detection in Ensemble Approaches for Workout Pose Recognition

Aman Rehman
Department of Computer Science and Engineering,
Delhi Technological University
Bawana Road, Delhi-110042, India
amanrehman18@gmail.com

Shailender Kumar
Department of Computer Science and Engineering,
Delhi Technological University
Bawana Road, Delhi-110042, India
shailenderkumar@dce.ac.in

***Abstract*—The identification and estimation of exercise poses has been a field of ongoing research in computer vision. Many deep learning architectures have demonstrated impressive performance, and as a result, much progress has been achieved in this area over the years. These developments highlight the enormous potential that exists at the nexus of fitness-related applications and computer vision, which applies to both images and videos. The complex task of workout image classification is addressed here using keypoint based classification, using YOLOv8 and MediaPipe to identify keypoints in the image and using those keypoints for classification. Then rigorous experiments were performed on a custom dataset designed to cover a wide range of poses—including differences in camera viewpoints, angles, and exercises. As part of this research, multiple algorithms were trained, including ensemble learning algorithms in an attempt to differentiate between the model performance of each algorithm. The effectiveness and limitations of different approaches have been measured in correctly identifying and classifying workout poses from 2D images. Analysis using multiple techniques is presented, where LightGBM with MediaPipe shows the best performance, achieving an accuracy of 95.95%.**

Keywords—deep learning, ensemble learning, keypoint detection, computer vision, pose detection

I. INTRODUCTION

Pose identification means determining the way a person positions their body when performing any activity. That position is then associated with that of a particular exercise, given the context learned from the image or video or real-time camera. Accurate identification of exercises is necessary for a variety of use cases, for example, supporting virtual personal trainers in guiding users through exercises with individualized recommendations, allows fitness tracking programs [1] to monitor workout routines and provide real-time feedback on form, and help with rehabilitation monitoring [2]. It also improves gesture detection in fitness games for immersive experiences, makes it easier to analyse athletes' actions for sports performance analysis [3], and can also connect with health monitoring devices to track physical activity levels and offer personalized recommendations.

Many exercise image recognition implementations use deep learning solutions, like CNNs, that is, Convolutional Neural Networks which are well-known for their effectiveness in classification and keypoint identification tasks [4]. However, the efficiency of these hinges significantly on the variety of the dataset. Ensuring that the model can represent the subtleties present in the given setting requires a strong dataset that covers a broad range of exercises, angles, and ambient conditions. Also, exercises are dynamic in nature. Further, many applications of exercise pose recognition may require fast inference and analysis, for example real-time analysis [5]. In such cases, the processing capacity of the underlying hardware is also an important consideration. This is because some people may prefer working out themselves with the help of a mobile application. Or they may be unable to hire a personal trainer. As a result, it is essential to find stability within accuracy of the model and computing resource limits because the application should be efficient for the given context.

In this paper, the process involves 2 approaches for keypoint detection by using cutting-edge methods, YOLOv8 (You Only Look Once) and MediaPipe. Both of these offer robust skeletal frameworks by obtaining keypoints from the images. Then that data is fed as input for training to various machine learning models like Decision Tree, Linear Support Vector Machine, Multi-Layer Perceptron, K-Nearest Neighbors and Gaussian Naive Bayes. Ensemble approaches like XGBoost, Bagging, LightGBM and Random Forest have also been used. Also to work with class imbalance due to some class having minority representation, SMOTE technique is applied. Then comparative evaluation is performed on various metrics with respect to all the models.

II. RELATED WORK

Pose identification and correction using machine learning methods has been extensively researched. It involves either entering exercises manually before performing them or automatically identifying and classifying user motion into an activity. There are studies that have used either images, or videos or both. In paper [6], the authors evaluated the motion quality of the workout performed by a user against the known correct motion. The authors used the HPTE dataset consisting of eight exercises performed by 5 different people from different backgrounds and angles. The proposed approach used PoseNet model in the backend which used two-dimensional data that has confidence linked to the body's important points and produces a heatmap for stance assessment. The system reached an overall accuracy of 80%. The user received feedback upon the completion of the task.

Bhat et al. [7] detected which joints need to be fixed during exercise by finding the overall joint angle distribution of the dataset and detecting poses in which a joint angle is an anomaly. The researchers created skeleton layouts using a variety of datasets based upon Yoga, Kungfu and Pilates. With the Kungfu-7 dataset, the DenseNet+Random Forest classifier delivered the best result, 86% validation accuracy.

979-8-3503-7131-4/24/$31.00 ©2024 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 24,2024 at 04:36:38 UTC from IEEE Xplore. Restrictions apply.

Vimala et al. [8] offered an artificial intelligence-based solution to follow patients' physical activity types. Two deep learning techniques were used: a recurrent neural network (RNN) and a deep feed-forward neural network (DNN). Two datasets, one with 14 daily life activities from wrist-worn sensors and the other with 10 activities from hip-worn sensors were used to evaluate model performances among volunteers. The RNN model performed better with 84.89% accuracy. The F1-score was 82.56%. This suggested the model's capacity to accurately understand as well as track the physical activities of patients.

Rangari et al. [9] focused on individual exercise pose recognition using 2D coordinates for body keypoints recorded by an RGB camera. 18 major body joints' coordinates were identified and used as the model's attributes for exercise classification. Human volunteers of various ages and heights were taken to create a benchmark dataset. When tested on that dataset, the model outperformed previous methods with an impressive accuracy of 97.01%, demonstrating the efficacy of their proposed strategy.

Joshi et al. [10] attempted to classify six different sports categories with respect to athlete's body movement. Inception V3 was used for derivation of feature and then neural networks for pose categorization. On comparison with other machine learning classifiers, it validated the framework's efficacy. The suggested method successfully detected and classified a variety of sports data, evident by its remarkable 96.64% average accuracy. Chen et al. [11], utilized computer vision techniques where the suggested system extracted the human body's contour, skeleton, dominating axes, and feature points to examine posture from front and side views. Then, directions for posture rectification were offered so that the posture could be altered.

In an another approach, Yubin et al. (2021) [12] proposed a Mediapipe based computer vision-based technique for rating yoga poses utilizing two datasets: Dataset A, which they took from Kaggle and Dataset B, which the researchers created. Using coarse as well as fine contrastive samples, the method produced the greatest accuracy for Dataset A, 0.8321.

Singh et al. [13] put forth a novel approach to position estimation and identification which did not require wearable technology. To reliably categorize different postures, two pre-trained models' performances were compared with those of a bespoke CNN called DeepPose, which was trained using both image and keypoint datasets. The outcome demonstrated that the suggested model performed better than the image dataset when using the keypoint dataset.

A self-prepared dataset as well as a publicly available one were used to demonstrate the findings of paper [14] for WorkoutNet, a deep learning network for exercise action identification. It surpassed models like as XceptionNet, VGG-19, and DenseNet121, with a validation accuracy of 92.75%. It was verified with F1-measure and confusion matrix to properly classify exercises into 10 groups. Faisal et al. [15] used a deep learning-based model that can recognize five distinct poses with significantly less data. The variability in the dataset is compensated by improving the diversity in the images. They evaluated performance of their model against many image classification models, including ResNet, InceptionResNet, and Xception. With their suggested architecture, each of the image's spatial and depth information is extracted and taken into account for additional calculations. According to the experiments, it attained 94.91% accuracy and 95.61% precision.

III. Proposed Approach

There are several steps in the procedure used for the assigned work. The proposed approach is demonstrated in Fig. **1**.

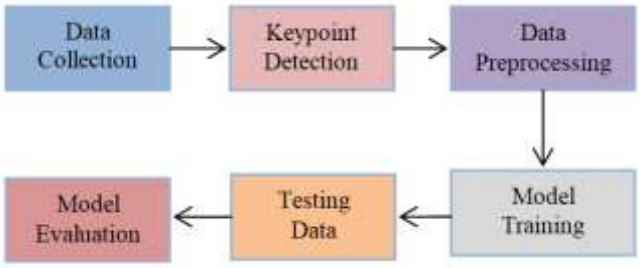


Fig. 1. Proposed Approach

A. Dataset Collection

Popular exercises were selected, and for each exercise, images were manually retrieved using Kaggle [16] as a guide and Google search engine. Consequently, the dataset has 1058 images, divided into 20 classes, each including 55–60 images. The twenty workout categories are as follows: bridges, bench press, bow, crane pose, crunches, cycling, cobra, child's pose, donkey kicks, downdog, deadlift, flutter kicks, pullup, plank, standing toe reach, shoulder press, squat, and triceps pushdown. A variety of stretches, bodyweight exercises, and weight training exercises are included in this set.

Tensorflow is then used to load images with a standard size of 224x224. Every image has a label based on the exercise being performed. Before loading, the dataset is shuffled to prevent imbalanced splits, following which a 80:20 split is made for training and testing respectively.

B. Keypoint Detection

YOLOv8 and MediaPipe algorithms have been used as the keypoint extraction algorithms. For YOLOv8, a grid is created using the input image, which gives cells. Now for each grid cell, bounding boxes and corresponding confidence ratings are predicted as part of a single-stage object recognition approach. The architecture usually comprises of many convolutional layers followed by upsampling and concatenation operations. Anchor boxes are used to increase localization accuracy, while feature fusion and skip connections are used to record spatial relationships.

The second method, MediaPipe extracts keypoints by first preprocessing the input image, then utilizing a convolutional neural network model trained for predicting the location of specific keypoints by detecting and localizing keypoints on the human body, including joints and landmarks. Features are extracted from the image at different scales, followed by regression techniques to predict the coordinates of each keypoint. Post-processing techniques may be applied to refine the results. That extracted keypoints act as the feature for each image. An example of keypoint detection using MediaPipe on a test image taken from the dataset [16] is shown in Fig. **2**.

C. Data Preprocessing

Preprocessing involves augmentations, which are required to increase the dataset's diversity to improve generalization. SMOTE (Synthetic Minority Over-sampling) Technique is a method frequently used in data augmentation to reduce class imbalance. By fabricating synthetic records for the minority class, it mitigates imbalance issues. First, a sample is selected at random from the minority class, and then its k-nearest neighbors are found. After calculating a

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 24,2024 at 04:36:38 UTC from IEEE Xplore. Restrictions apply.

vector between the current data point and chosen neighbor, a random number between 0 and 1 is multiplied. For every image in the dataset, this procedure is repeated. The labels were categorical in

Fig. 2. Keypoint detection in an exercise image

nature. They are encoded into numerical values using LabelEncoder.

D. Model Training

Thereon multiple algorithms have been trained on both the above skeletal data to compare the performance between them. For each algorithm, GridSearchCV was used to find the best hyperparameters. It is a method for systematically tuning hyperparameters of machine learning models. It works by searching through a specified list of parameters and also utilizes cross-validation to evaluate each combination's performance. The following models are used.

GaussianNB: It is a Bayes' theorem based probabilistic classifier, assuming normal distribution of features. It may be used to categorize keypoints according to probability distribution to differentiate between exercise postures.

K-nearest neighbors (KNN): KNN uses majority class voting to classify items using its K nearest neighbors. By calculating the distances between the keypoints of labeled instances, it can determine which class label the new instance most closely resembles [17].

Linear support vector machines: Linear SVM is utilized to effectively discriminate between postures by classifying them into distinct exercise categories based on spatial arrangement.

Decision tree: It is a tree structure with leaf nodes representing class labels and branches are the decision rules. Exercise posture categorization can be facilitated using this technique by generating hierarchical judgments based on the attributes of keypoint [18].

Random forest: Here, multiple decision trees are constructed. These then provide us with classification using the mode of the identifications for each tree. The performance hinges on the quality of the features found.

LightGBM: It uses a leaf-wise splitting technique. It divides the feature values into bins and then creates histograms on it due to which it effectively manages large datasets while reducing processing cost [19].

Bagging: With the help of voting, it combines the results of multiple classifiers which have been previously trained on varying bootstrap samples of the training data.

XGBoost: Based on gradient boosting, it is a boosted ensemble of decision trees for keypoint identification, XGBoost can efficiently capture intricate interactions between keypoints [20].

Multi layer perceptron (MLP): It is a neural network made up of several layers of neurons and has connections between every neuron in layers above and below.

E. Model Evaluation

The test set was used for evaluation of the models. On each model, several techniques were applied for a thorough analysis.

Then their test accuracy, precision and recall were determined.

A comparative analysis was conducted in order to further understand each models' performance with respect to the keypoint detection techniques used. The model that performed the best and had the highest test accuracy was then selected, and its weights were saved.

IV. Experimental Analysis

A. Requirement and Specification

Python version 3.10 was used to conduct the study. Hardware requirements include an Nvidia RTX 3080 GPU, a 4-core Intel i7 processor, 32GB DDR4 RAM, and the use of Google Colab's GPU for faster processing.

B. Experimental Details and Results

YOLOv8 predicted a total of 32 keypoints in this experiment while MediaPipe produced 33. TABLE I listed the best parameters found using Grid Search with respect to each machine learning and ensemble technique upon training on each skeletal data.

According to Table II, classifiers that use MediaPipe keypoints as opposed to YOLOv8 achieved higher accuracies. For instance, Linear SVM with MediaPipe shows accuracy increase from 69.88% to 90.90%, while GaussianNB's accuracy increases from 56.79% to 88.88% as well. LightGBM performed best with MediaPipe with 95.95%. Random forest closely follows it with 94.94%. Other ensemble techniques like Bagging also shows higher accuracy rates. Additionally, Multi-Layer Perceptron shows a notable improvement in accuracy employing MediaPipe keypoints, going from 81.89% to 91.41%.

For YOLOv8, Bagging showed the best performance of 82%. Further, in contrast to MediaPipe, Decision Trees and KNN show comparatively poorer accuracy rates when using YOLOv8. These results highlight how important feature extraction methods are to improving classifier performance.

TABLE I BEST HYPERPARAMETERS FOUND USING GRIDSEARCHCV

Model	Parameter	YOLOv8	MediaPipe
Decision Tree	criterion	entropy	entropy
	max_depth	none	none
	min_samples_split	2	1
	min_samples_leaf	2	2
	splitter	best	best
Linear SVM	c	1	1
	loss	hinge	hinge
	penalty	l1	l2
Multi Layer Perceptron	activation	relu	relu
	alpha	0.0001	0.0001
KNN	n_neighbors	3	3
	weights	distance	distance
	p	2	2
XGBoost	learning_rate	0.001	0.001
	n_estimators	128	100
	max_depth	5	3
	gamma	0.1	0.2

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 24,2024 at 04:36:38 UTC from IEEE Xplore. Restrictions apply.

Bagging	n_estimators	128	100
LightGBM	learning_rate	0.003	0.001
	n_estimators	100	100
	max_depth	3	5
Random Forest	n_estimators	128	100
	max_depth	none	5
	min_samples_split	2	5
	min_samples_leaf	1	1

Compared to using just one model, ensemble algorithms integrate several models to get more accurate predictions. They increase strengths and minimize weaknesses by utilizing the diversity of several models. This is reflected in the results with all ensembles showing better performance.

Overall it can be said that the combination of MediaPipe with ensemble algorithms show consistently good performance when compared with other combinations.

TABLE II TEST ACCURACIES OF ALL CLASSIFIERS

Individual classifier	Accuracy (YOLOv8)	Accuracy (MediaPipe)
Decision Trees	50.00%	84.34%
Linear SVM	69.88%	90.90%
Multi-Layer Perceptron	81.89%	91.41%
KNN	66.68%	70.20%
GaussianNB	56.79%	88.88%
XGBoost	73.00%	91.91%
Bagging	**82.00%**	93.43%
LightGBM	80.00%	**95.95%**
Random Forest	78.00%	94.94%

To obtain more fine-grained understanding of the performance, precision, recall and F1 score are shown in Table III. Precision tells us out of all the positively predicted occurrences, what is the percentage of correctly predicted positive instances. Recall gives us the percentage of accurately predicted positive cases among all true positives. It tells us how well the model recognizes positive cases. F1 score gives us a single metric for model evaluation which is defined as the harmonic mean of precision and recall.

TABLE III PRECISION, RECALL AND F1 SCORE FOR MEDIAPIPE METHODS

Algorithm	Precision	Recall	F1 score
Decision Tree	0.83	0.85	0.83
Linear SVM	0.89	0.87	0.88
Multi-Layer Perceptron	0.89	0.90	0.88
KNN	0.67	0.64	0.64
GaussianNB	0.87	0.88	0.87
XGBoost	0.96	0.95	0.95
Bagging	0.97	0.95	0.96
LightGBM	0.97	0.96	0.96
Random Forest	0.95	0.94	0.95

Accurate classification and a reduction in false positives using MediaPipe keypoints is demonstrated by most ensembles, which routinely attain precision and recall above 0.95. When it comes to reliably classifying images based on MediaPipe keypoints, Linear SVM and Multi-Layer Perceptrons also fare well, scoring around 0.89. Lower ratings for KNN, however, point to possible false positives and limitations in its classification. The findings highlight the significance of choosing the right technique with MediaPipe-based features for classification tasks.

The confusion matrix in Fig, provides a breakdown of exercise classifications. The accurate pose predictions are demonstrated via diagonal elements. The regions that are misclassified and hence need improvement are the off-diagonal elements. For every exercise, it displays the wrong class that the model predicts. From the confusion matrix, the system demonstrates 100% accuracy for a number of exercise categories, namely bench press, cobra, child's pose, downdog, deadlift, flutter kicks, leg extensions, lat pulldowns, plank, standing toe reach, shoulder press and triceps pushdown.

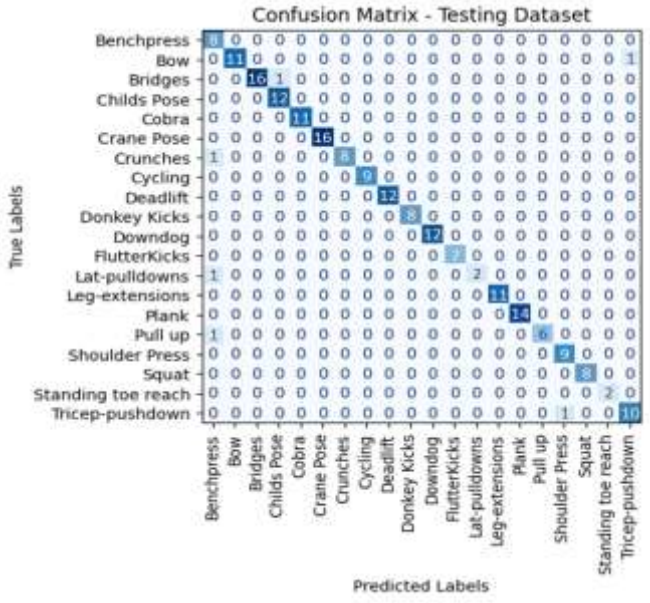


Fig. 3 Confusion matrix of LightGBM model with MediaPipe landmarks

Table IV shows the pose-wise accuracy for MediaPipe+LightGBM model. This is confirmed by its high precision and recall.

TABLE IV ACCURACY OF EACH WORKOUT CLASS FOR LIGHTGBM MODEL (WITH MEDIAPIPE KEYPOINTS)

Pose	Accuracy
Bridges	94.11%
Bench press	100.00%
Bow	91.66%
Cobra	100.00%
Child's pose	100.00%
Crane pose	93.75%
Cycling	88.88%
Crunches	88.88%
Downdog	100.00%
Donkey kicks	87.50%
Deadlift	100.00%
Flutter kicks	100.00%
Leg extensions	100.00%
Lat pulldowns	100.00%
Pullups	85.71%
Plank	100.00%
Squat	87.50%
Standing toe reach	100.00%
Shoulder press	100.0%
Tricep pushdown	100.00%

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 24,2024 at 04:36:38 UTC from IEEE Xplore. Restrictions apply.

Fig. 3 displays the learning curve for the LightGBM model with the number of training instances on the x-axis and the cross-validation score on the y-axis. The cross-validation score is shown by the green line, which initially increases rapidly from around 0.6 to 0.95 as more training samples are added. This indicates that the performance of the model is improving and it is becoming more proficient at extracting features from the data. But after a certain point, the score begins to plateau, as seen from the slow increase in score from 0.95 to 0.96, indicating that the model's performance may not significantly enhance further by adding more training data.

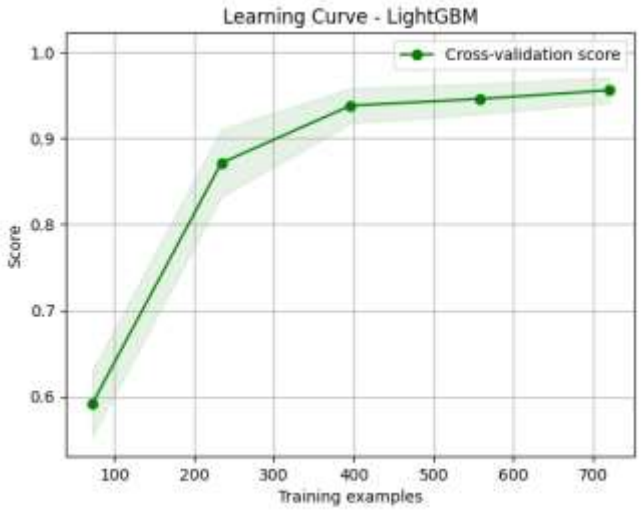


Fig. 3 Cross-validation score vs number of samples

The Fig. 4 shows the AUC-ROC curve for each class in the MediaPipe keypoints based LightGBM model. The ROC curve displays the true positive rate versus the false positive rate. The curve closest to the top-left corner represents the Benchpress class, indicating perfect classification performance. Classes like Bridges, Cobra and many others also have ROC curves hugging the top-left corner, suggesting excellent discrimination ability.

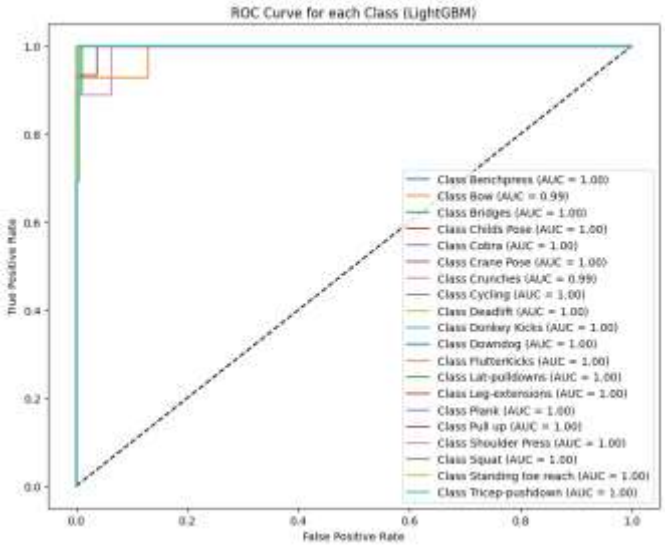


Fig. 4 AUC-ROC curve for each class Mediapipe+LightGBM)

C. Output

In Fig. 5, photos from testing dataset [16] of various exercise positions are taken, and the true and predicted labels are displayed. Also, the predicted keypoints on the poses are also showcased.

V. Conclusion

In this research, the domain of action recognition was explored by utilizing multiple techniques using multiple algorithms, using keypoint detection via YOLOv8 and MediaPipe. During training, data augmentation techniques were used to correct class imbalance and improve resilience. The classification step was carried out after data augmentation and keypoint extraction. MediaPipe with ensemble algorithms consistently showcased impressive performance. However YOLOv8 was found to be inferior to MediaPipe in all the tests, even though the number of keypoints that were found to be nearly the same for both. This also shows the capability of

Fig. 5 True vs Predicted pose of a few test samples

ensemble algorithms in how the different models combine their learned features to improve performance.

A. Future Work

Future work will investigate the use of transformer-based models for vision, namely Vision Transformers which is a new architecture proposed for computer vision tasks. Vision transformers [21] have performed extremely well on a number of computer vision tasks and have also outperformed many state-of-the-art benchmarks previously held by deep convolutional neural networks. Further, this domain can benefit from the inclusion of explainable techniques that give clarity about the output generated. For example, if an exercise was classified incorrectly, then which features caused that misclassification. Incorporating explainability into exercise feedback systems can also provide real-time explanation about the pose performance. This will enable researchers to obtain insights into the model understanding and improve performance for the future applications.

REFERENCES

[1]. Y. Jain, D. Chowdhury, and M. Chattopadhyay, 'Machine Learning Based Fitness Tracker Platform Using MEMS Accelerometer', in *2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*, Dec. 2017, pp. 1–5. doi: 10.1109/ICCECE.2017.8526202.

[2]. W. Zhang, C. Su, and C. He, 'Rehabilitation Exercise Recognition and Evaluation Based on Smart Sensors With Deep Learning Framework', *IEEE Access*, vol. 8, pp. 77561–77571, 2020, doi: 10.1109/ACCESS.2020.2989128.

[3]. R. Ji, 'Research on Basketball Shooting Action Based on Image Feature Extraction and Machine Learning', *IEEE Access*, vol. 8, pp. 138743–138751, 2020, doi: 10.1109/ACCESS.2020.3012456.

[4]. A. Newell, K. Yang, and J. Deng, 'Stacked Hourglass Networks for Human Pose Estimation', in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 483–499. doi: 10.1007/978-3-319-46484-8_29.

[5]. A. Moran, B. Gebka, J. Goldshteyn, A. Beyer, N. Johnson, and A. Neuwirth, 'Muscle Vision: Real Time Keypoint Based Pose

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 24,2024 at 04:36:38 UTC from IEEE Xplore. Restrictions apply.

Classification of Physical Exercises'. arXiv, Mar. 22, 2022. doi: 10.48550/arXiv.2203.12111.
[6]. H. Rookba *et al.*, 'Automatic Feedback For Physiotherapy Exercises Based On PoseNet', *FCI-H Inform. Bull.*, vol. 2, no. 2, pp. 10–14, Oct. 2020, doi: 10.21608/fcihib.2020.116046.
[7]. B. Dittakavi *et al.*, 'Pose Tutor: An Explainable System for Pose Correction in the Wild', in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2022, pp. 3539–3548. doi: 10.1109/CVPRW56347.2022.00398.
[8]. V. Nunavath *et al.*, 'Deep Learning for Classifying Physical Activities from Accelerometer Data', *Sensors*, vol. 21, no. 16, p. 5564, Aug. 2021, doi: 10.3390/s21165564.
[9]. T. Rangari, S. Kumar, P. P. Roy, D. P. Dogra, and B.-G. Kim, 'Video based exercise recognition and correct pose detection', *Multimed. Tools Appl.*, vol. 81, no. 21, pp. 30267–30282, Sep. 2022, doi: 10.1007/s11042-022-12299-z.
[10]. K. Joshi, V. Tripathi, C. Bose, and C. Bhardwaj, 'Robust Sports Image Classification Using InceptionV3 and Neural Networks', *Procedia Comput. Sci.*, vol. 167, pp. 2374–2381, 2020, doi: 10.1016/j.procs.2020.03.290.
[11]. H.-T. Chen, Y.-Z. He, and C.-C. Hsu, 'Computer-assisted yoga training system', *Multimed. Tools Appl.*, vol. 77, no. 18, pp. 23969–23991, Sep. 2018, doi: 10.1007/s11042-018-5721-2.
[12]. Y. Wu *et al.*, 'A Computer Vision-Based Yoga Pose Grading Approach Using Contrastive Skeleton Feature Representations', *Healthcare*, vol. 10, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/healthcare10010036.
[13]. S. Manvendra, A. A. M. Sarfaraj, and C. G. Mahesh, 'DeepPose: An Integrated Deep Learning Model for Posture Detection Using Image and Skeletal Data', *IEEE Conf. Proc.*, vol. 2023, no. ICCCNT, pp. 1–7, 2023.
[14]. A. Dey, A. Dutta, and S. Biswas, 'WorkoutNet: A Deep Learning Model for the Recognition of Workout Actions from Still Images', in *2023 3rd International Conference on Intelligent Technologies (CONIT)*, Jun. 2023, pp. 1–8. doi: 10.1109/CONIT59222.2023.10205926.
[15]. F. B. Ashraf, M. U. Islam, M. R. Kabir, and J. Uddin, 'YoNet: A Neural Network for Yoga Pose Classification', *SN Comput. Sci.*, vol. 4, no. 2, p. 198, Feb. 2023, doi: 10.1007/s42979-022-01618-8.
[16]. 'Workout/Exercise Images'. Accessed: Dec. 19, 2023. [Online]. Available: https://www.kaggle.com/datasets/hasyimabdillah/workoutexercises-images
[17]. D. Shah, V. Rautela, C. Sharma, and A. Florence A, 'Yoga Pose Detection Using Posenet and k-NN', in *2021 International Conference on Computing, Communication and Green Engineering (CCGE)*, Pune, India: IEEE, Sep. 2021, pp. 1–4. doi: 10.1109/CCGE50943.2021.9776451.
[18]. A. J. Torres, C. Silubrico, D. Torralba, and J. P. Tomas, 'Detection of Proper Form on Upper Limb Strength Training Using Extremely Randomized Trees for Joint Positions', in *Proceedings of the 2nd International Conference on Computing and Big Data*, in ICCBD 2019. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 111–115. doi: 10.1145/3366650.3366680.
[19]. K. Oi, Y. Nakamura, Y. Matsuda, M. Fujimoto, and K. Yasumoto, 'Short Stick Exercise Tracking System for Elderly Rehabilitation using IMU Sensor', in *2022 2nd International Workshop on Cyber-Physical-Human System Design and Implementation (CPHS)*, May 2022, pp. 13–18. doi: 10.1109/CPHS56133.2022.9804564.
[20]. J. Guo *et al.*, 'An XGBoost-based physical fitness evaluation model using advanced feature selection and Bayesian hyper-parameter optimization for wearable running monitoring', *Comput. Netw.*, vol. 151, pp. 166–180, Mar. 2019, doi: 10.1016/j.comnet.2019.01.026.
[21]. W. Chen *et al.*, 'A Simple Single-Scale Vision Transformer for Object Localization and Instance Segmentation'. arXiv, Oct. 01, 2022. doi: 10.48550/arXiv.2112.09747.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 24,2024 at 04:36:38 UTC from IEEE Xplore. Restrictions apply.

Signal, Image and Video Processing
https://doi.org/10.1007/s11760-024-03275-1

ORIGINAL PAPER

Link chain driven reversible data hiding in encrypted images for high payload

Ankur[1] · **Rajeev Kumar**[1] · **Ajay K. Sharma**[2]

Received: 12 April 2024 / Revised: 30 April 2024 / Accepted: 5 May 2024
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2024

Abstract
In the last decade, data hiding in encrypted image has got a lot of attention within the security and privacy researcher's communities. This attention has led to the development of various reversible data hiding techniques for encrypted images (RDHEI) which enables separable lossless recovery of cover media and extraction of the secret message, at the receiving end. In this paper, a link chain driven (LCD) RDHEI algorithm for providing high embedding capacity, is proposed. The LCD-RDHEI exploits neighbourhood similarity of image pixels and clubs them in variable size link chains. The pixels in each link chain are estimated either by their average or using two quantization values based on the size of link chains. This helps in reserving a big room inside the image which can be used to fill in the secret information. Additionally, the pixels which are not part of any of the link chains are also utilized for embedding the secret information by adaptive multi MSB prediction and Huffman Coding method. Experimental results show that the LCD-RDHEI provides significantly greater embedding capacity than the existing and related modern data hiding method for encrypted images.

Keywords Link chain · Reversible data hiding · Encrypted images · RDHEI · Huffman · AMBTC

1 Introduction

In recent years, there has been an exponential growth in the size of cloud storage as people are using cloud storage for saving their data that is usually in the form of images, audios, videos, documents, etc., Behind the surge in the usage of the cloud storage, it has been the cloud storage's advantage such as usability, accessibility, cost-efficiency, convenience, scalability, etc., over the traditional way of storage, however, at the same time, the cloud storage lacks total control on security and privacy[1]. For this, data hiding appears one of the most promising options in order to obtain the basic goals of security i.e., confidentiality, authentication, and integrity of cloud users [1, 2]. Data hiding makes immersion of secret data in a multimedia cover possible and allows flawless extraction of the same during the process of decoding [3]. That's why, the data hiding (in short DH) finds its applications in domains such as military, medicine, judiciary and mass media where confidentiality of data in transit, are of utmost importance [4, 5].

Reversible data hiding (RDH) is an extension of DH that ensures the rebuilt of multimedia cover in the process of extraction of the immersed secret data which was not guaranteed in the case of traditional data hiding [6]. With the advent of cloud technology and the flow of large amounts of data that is usually in the form of images over the Internet, RDH finds images as the potential covers. Additionally, the images in this context may also be encrypted to create a better level of security. Therefore, to provide the required security, reversible data hiding in encrypted images (in short RDHEI) have been introduced which immerses data in encrypted images [7, 8]. RDHEI allows the data owner to hide their data in a cover image even without knowing the image content. The primary objective of the RDHEI is to achieve a higher embedding capacity and low distortion in the image with high security of data. However, the problem in achieving higher embedding capacity and low distortion at the same time, is their trade-off, in other words, the mechanism of data immersion in an image brings in a distortion in the original image cover. So, the larger the data is immersed the bigger

✉ Rajeev Kumar
rajeevkumar@dtu.ac.in

[1] Department of Computer Science and Engineering, Delhi Technological University, Delhi 110042, India

[2] Department of Computer Science and Engineering, National Institute of Technology, Delhi 110036, India

Published online: 28 June 2024

will be the distortion. In the proposed work, an effort is put to optimize this traditional trade-off for improved performance.

Over the last few years, various methods have evolved in RDHEI [7–30]. These methods take into consideration numerous techniques such as data compression, spatial correlation, histogram frequency and other encryption techniques for creating the room in the image to embed the data. However, as far creating the room inside the image is concerned, there are majorly two categories namely, Vacating room after encryption (VRAE) and Reserved room before encryption (RRBE). The VRAE techniques [7–14] usually create a room for embedding the secret information after the image encryption. In other words, VRAE does not demand the image owner to pre-process his/her original image for creating space for data immersion. VRAE techniques usually have smaller room for data immersion in the multimedia cover as the encrypted image has its level of entropy at maximum and thus leaves minimum spatial correlation. Various VRAE based data hiding works have been discussed, however, most of the work has limited embedding capacity. In contrast, the RRBE technique as its name suggests, reserves room before encryption of image which demands the image owner to have an idea regarding usage of his image as a multimedia cover so that he can leave room for data immersion within the image. Thus, grants a higher amount of secret data to be immersed in them.

Various RRBE based methods have already been proposed in recent years [15–30]. One such and one of earliest as well is discussed by Ma et al. [15] in 2013. In this method, a room is created inside the image using histogram shifting. Here, the shifting of the first three Least significant bits (LSBs) is done from one region to another region. This method does not guarantee a perfect reconstruction of the original image. Moreover, its embedding capacity is very low. Zhang et al. [16] discussed a data hiding method for cipher images obtained after encrypting by public-key cryptosystems with probabilistic and homomorphic properties. They embed the secret information into LSBs of the image which causes very low distortion and in turn allows lossless recovery of the original image. Cao et al. introduced a RDHEI method by borrowing sparse coding from the sparse representation technique. The sparse coding allows the hider to exploit the local correlation and creates a larger room inside the image for embedding the secret information. Yi et al. [18] discussed an RDH method for encrypted images using binary block embedding (BBE). The method introduces the usage of the most significant bits (MSB) planes for embedding the data and provides a significant increase in the data hiding capacity over the erstwhile RDHEI methods. Additionally, a bit level scrambling process is also employed to fully scatter the data inside the image for thwarting brute-force and similar attacks. Puteaux et al. [19] extended the work of MSB based embedding in the RDHEI domain by introducing two hiding methods. The first method basically replaces all the MSB bits with data and obtains an embedding capacity of 1 bpp (bit per pixel) with the resultant image being quite similar to the original image. The second method replaces only suitable MSB bits and to achieve the identical image as the embedded data is a little less than 1 BPP. Puyang et al. [20] extended the work of [19] by exploiting the neighbourhood correlation and making use of two MSBs for data embedding. Chen et al. [21] discussed a high capacity RDHEI method based on extended run-length coding and block-based MSB plane re-ordering. The run-length coding has been used to compact the MSB planes of images for creating large room and in turn obtaining the higher embedding capacity.

Shiu et al. [22] discussed a RDHEI method which make use of two compression techniques i.e., interpolative AMBTC and Huffman coding. It first vacates the room using interpolative AMBTC within a block and generates a code, and then reduces the length of code by Huffman encoding. Thus, a room is created inside the blocks which can be used for embedding the secret information. In 2020, Yin et al. [23] introduced another extension of [19] by multi-MSB prediction and Huffman encoding. The method makes use of prediction strategy for predicting the pixel and compares the original pixel value with the predicted one in bitwise manner. For the error encoding, Huffman method has been used which helps in optimally representing the errors and save the space for creating a large room. Thus, large embedding capacity is achieved. In the prediction domain, another method which is based on parametric binary tree levelling (PBTL), was introduced by Wu et al. [24]. The PBTL method here divides the image into blocks and make use of median edge detector (MED) predictor for prediction. The errors here are encoded using PBTL method. The PBTL method provides good embedding capacity. Xiao et al. [25] use compressive sensing for increasing the data hiding capacity. However, the method does not provide lossless recovery of the cover image but the immersion capacity was very large, moreover high and constant PSNR was achieved. Weng et al. [26] discussed an RRBE-RDHEI method by re-arranging bits of 5 MSB-planes to achieve a highly correlated bitstream. Afterwards, the resulting bitstream is compressed using a group classification encoding and the created space after compression is used for data embedding. Another method to create a larger space inside the encrypted image was introduced by Gao et al. [27]. The introduced method makes use of difference in prediction errors and creates a label map which is then divided into groups. The groups are then encoded to create space for embedding. Thus, the method has been able to achieve more than 2bpp embedding capacity.

In 2023, Li et al. [28] proposed a method predicting the MSB of each pixel, similar to Puteaux et al. [19]. Post MSB prediction, embeddable bits are highlighted on an error map, followed by the division of bit planes, executing a cycle

exclusive OR operation. This intricate process ensures a smooth transition of anticipated bits to the LSB plane while preserving the high-bit plane's integrity. Consequently, the method is constrained to achieve embedding rate (ER) of approximately 1 bpp. Qin et al. [29] devised a method for enhanced ER using grey transformation, compressive sensing (CS), and intra-block difference coding. Initially, grey transformation is employed to prioritize important bits. Then, the matrix is blocked, and CS-based reconstruction is utilized for optimized image data representation. Further refinement is achieved via intra-block difference coding to extract block correlations. Despite achieving high ER, this method also leads to considerable distortion in the reconstructed image. To amplify the EC, Liu et al. [30] put forth a method combining arithmetic and Huffman encoding. The process begins by partitioning the image into bit-planes. Each bit-plane undergoes arithmetic encoding, followed by further compression through Huffman encoding. Additionally, the methodology incorporates scrambling to encrypt overhead, necessitating an extra key. Although the approach ensures lossless recovery of image and data, it lacks separability. Similarly, Gao et al. [31] proposed a hierarchical scheme that categorizes image pixels into multiple linear regressions (MLR) predictable and non-MLR predictable sets. The MLR-predictable pixels generate two label maps using adaptive adjustment of predicted data and multiple bits prediction, while the non-MLR pixels use simple one-bit prediction for a label map. By compressing and embedding these label maps into the encrypted image, ER of 2.99–3.88 bpp is achieved across datasets. Fu et al. [32] presented a new RDHEI method using "L"-shaped block embedding (LBE) and improved binary-block embedding, which fully utilize the Laplacian-like distribution of prediction errors from pixel prediction. This approach adaptively selects embedding methods based on net capacities, therefore significantly enhancing embedding capacity. However, both the methods [31, 32] lack in adaptability of increasing the EC at the cost of quality and vice-versa.

Based on the analysis of the related literature, it has been observed that the high embedding capacity is usually achieved by block wise prediction-based methods as the prediction errors can be encoded to create space which can be later used for embedding. However, due to the fixed size blocks, the complete neighbourhood correlation has not been fully utilized which leads to non-accurate prediction sometimes. Moreover, there is a need of the method which can provide flexible embedding capacity at the cost of PSNR. To address these problem, a new high-capacity data hiding method which tries to exploit the neighbourhood correlation, is proposed. The proposed method basically generates variable size link chains of highly correlated pixels and represents those pixels by one or two values i.e., their average or two quantization levels, respectively, based on the size of link chains. Thus, a large room inside the image is created which is used to embed the secret information. The main contribution of this work is as follows:

- The proposed work introduces a novel high-capacity data hiding method that leverages the similarity between neighboring pixels to create variable-length chains of connected pixels. These chains are then used to create space within the image, which can be utilized to embed secret information.
- The method involves estimating the values of pixels within each chain using either their average value or two quantization values, depending on the size of the chain. This estimation process helps in creating a big room within the image for hiding the secret data.
- The experimental results demonstrate that the proposed LCD-RDHEI method offers greater embedding capacity compared to the existing and related data hiding methods, while ensuring that the original image can be recovered after data extraction.
- The aforementioned capabilities of the proposed LCD-RDHEI method makes it ideal for cloud applications where sensitive data need to be stored within encrypted multimedia files for ensuring end-to-end data confidentiality, integrity while reaping the benefits of the accessibility and the limited cost.

In the next section, review of some of the most closely related data hiding methods in encrypted domain is provided.

2 Review of related work

In this section, review of the working of two most closely related works i.e., Yin et al. [23] and Shiu et al. [22] in the field of RDHEI, is presented. All of these aforementioned works belong to RRBE category and provide good embedding capacity. In the presented review, the complete working process of Adaptive Multi-MSB Prediction and Huffman Coding (AM2PHC) [23] followed by Interpolative Absolute Moment Block Truncation Coding (IAMBTC) based RDHEI methods, is precisely explained as follows.

2.1 Adaptive multi-MSB prediction and Huffman coding (AM2PHC) [23]

Yin et al. introduced a new RRBE method based on adaptive multi-MSB prediction and Huffman coding (AM2PHC) for reversible data hiding in encrypted images (RDHEI). AM2PHC-RDHEI is a completely lossless method which makes use of prediction strategy for reserving the room inside the image and provides high embedding capacity. For this, Yin et al. employs the median edge detector (MED) predictor used in the LOCO-I algorithm of the JPEG-LS

compression for prediction of the pixel values of the original image (I) of size M×N pixels. The pixel located at $(i,\ j)$ is predicted using MED predictor based on three closest causal pixels (i.e., $X = I(i-1,\ j)$, $Y = I(i,\ j-1)$ and $Z = I(i-1,\ j-1)$) as follow:

$$p(i,\ j) = \begin{cases} \max(X,\ Y) & if\ Z \leq min(X,\ Y) \\ \min(X,\ Y) & if\ Z \geq max(X,\ Y) \\ X + Y - Z & otherwise \end{cases} \quad (1)$$

where $2 \leq i \leq M, 2 \leq j \leq N$ and $p(i,\ j)$ is the predicted value of $I(i,\ j)$. Next, both the original and predicted pixel values are converted into 8-bits binary representation and compared in bitwise manner from MSB to LSB until a non-match bit is found. Then, a label that is equal to the number of matching bits (till the first non-matching) is created for the pixel. Therefore, the label value for embeddable pixels can vary from 0 to 8, where 0 represents the first MSB itself is different and 8 represents all the bits are matching, thus, total 9 values. The process of prediction and label generation is repeated for the entire image to get a label map, which contains label (t) for each pixel of the image. This label is used to embed $t + 1$ bits in that pixel. The first pixel, on the other hand, is a reference pixel that cannot be used for data embedding. The generated label map is then encoded using Huffman method which generally results into the following Huffman code i.e., {00, 01, 100, 101, 1100, 1101, 1110, 11110, 11111} for the label map values based on their frequency. Usually, the least frequent value is represented by maximum bits and vice-versa. Thus, an encoded label map (ELM) in the form of binary stream is obtained. Afterwards, the original image (I) is encrypted by performing a bit-wise exclusive-or operation on every bit of the original image using an encryption key. The encryption process is same as the one to be discussed in the proposed work. Next, the ELM is embedded in the encrypted image to generate space inside the image. Before embedding the secret data, Huffman coding rules (HCR), length of ELM and other partial auxiliary information which includes ELM and original reference pixel's values is embedded. This embedding is done in the certain pixels (also knowns as reference pixels) of the first row and the first columns by replacing their values. The partial embedding of auxiliary is done because it is not possible to replace all reference pixels in their original locations due to the size, separability and reversibility constraints. The value of t is used to guide the embedding process which basically enables embedding of $t + 1$ bits from MSB onwards except for the case of $t = 8$, where all the bits of the pixel can be embedded. After embedding of HCR and partial auxiliary information, the embedding of remaining auxiliary information and bits of secret data is done by data hider using the same embedding process defined above and final marked encrypted image is obtained.

The marked encrypted image is obtained by the receiver who can the extract the hidden information and restore the image based on the availability of only data key, only encryption key or both. In all three cases, first the HCR and auxiliary information are extracted. If the data receiver has only the data key, the encrypted data is extracted from the similar position as embedded during hiding, and then decrypted to obtain lossless data. In case the receiver has only encryption key, after extracting the HCR, $LELM$ and reference pixels, the image is decrypted using the same process of encryption. Although, the image reconstruction process is done in raster scan manner and pixel values are predicted using Eq. (1), with the exception of the first row and column. And the appropriate predicted value bits are used to get the original pixel values. Following the completion of this replacement, a lossless image is obtained. In case both hiding and encryption keys are available with the same receiver, the above defined process of extraction and recovery can be executed in a sequential manner to get both hidden data and recovered original image.

2.2 Shiu et al.'s IAMBTC [22]

Shiu et al. [22] discussed an interpolative absolute moment block truncation coding [33] (IAMBTC) based high-payload RDH method for encrypted images. IAMBTC technique is an extended version of Absolute Moment Block Truncation Coding (AMBTC) approach given by Lema et. al. [34] in 1984. Shiu et al.'s method first divides the image into 4×4 blocks into two categories namely hiding and saving based on their variance. If the variance is less than a user-defined threshold then the block is a hiding block, otherwise, it is a saving block. In Fig. 1, block (a) and (k) represents hiding and saving blocks respectively. Next, the saving blocks are marked at the 4th pixel LSB to discriminate them from the hiding block as shown in block (l) and the hiding blocks are encoded using IAMBTC. For this, an average (h) of each hiding block is calculated using Eq. (2).

$$h = \left\lfloor \frac{1}{n} \sum_{i=1}^{n} x_i \right\rfloor \quad (2)$$

where n represents the number of pixels in a block and x_i is the intensity of i^{th} pixel. Next, two quantization values (H^l and H^h) are computed using Eqs. (3) and (4), respectively.

$$H^l = \approx \left[\frac{1}{p} \sum_{i=1}^{n} x_i \right] \quad if\ x_i \leq h \quad (3)$$

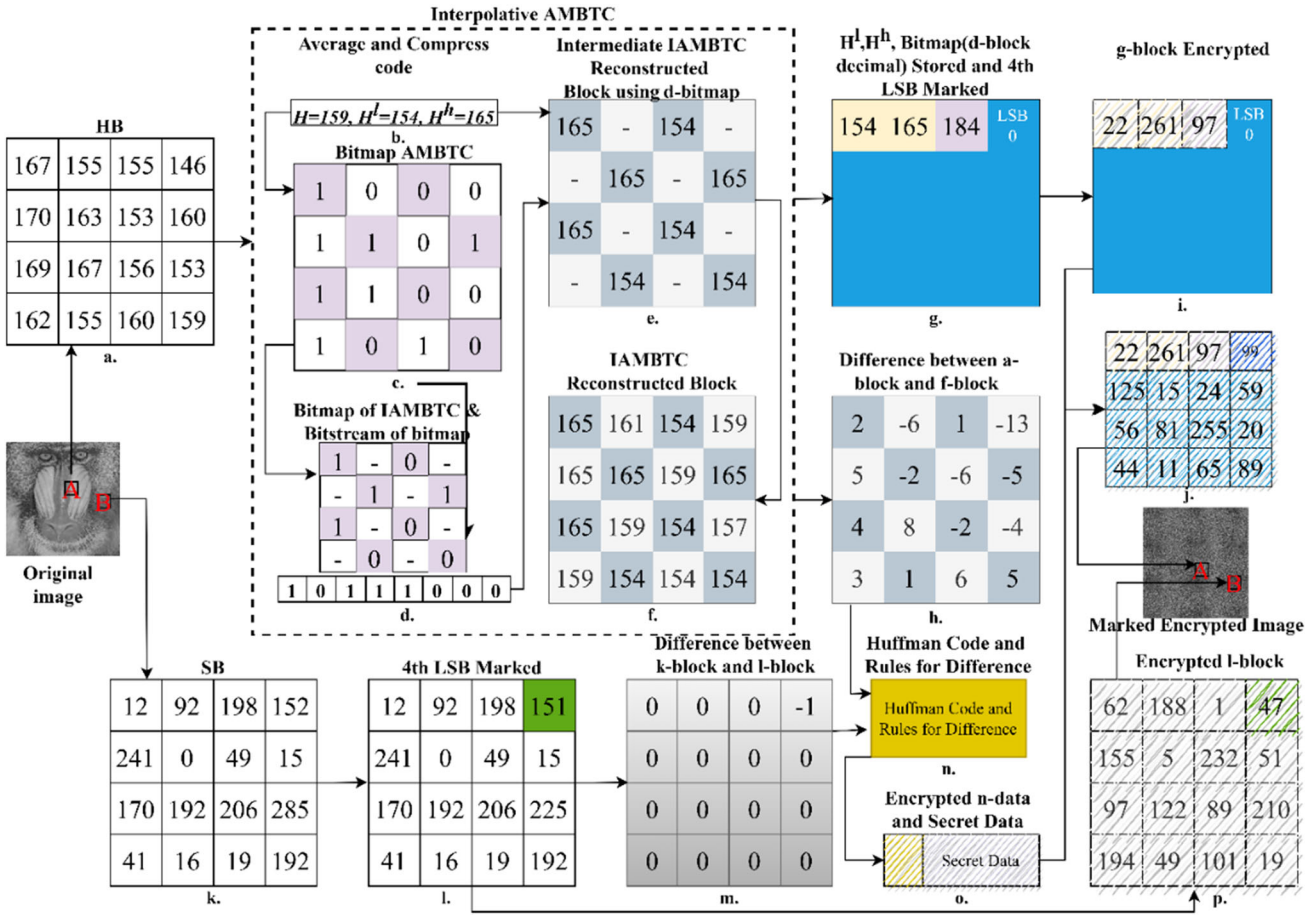


Fig. 1 An example of pre-processing, marking, and data hiding of the IAMBTC-RDHEI method

$$H^h = \approx \left[\frac{1}{n-p} \sum_{i=1}^{n} x_i \right] \quad if x_i > h \tag{4}$$

where p denotes the count of pixels having value less than h and '$=\approx$' represents assignment to the nearest whole number. H^l and H^h represent low and high quantization values, respectively, as shown in block (b) of Fig. 1. Additionally, a bitmap is also created which represents each pixel of the block by a bit which is '1' if the pixel value is greater than h otherwise it is '0', as shown in block (c) of Fig. 1. To further reduce the size, every alternate (even) bit of the bitmap is dropped as shown in block (d) of Fig. 1 to get a partial bitmap $\left(B^M\right)$. Thus, the hiding block is represented by trio i.e., $\left\{H^l,\ H^h,\ B^M\right\}$. To reconstruct the block, odd numbered pixels can be reconstructed using following Eq. (5), as shown in block (e) of Fig. 3.

$$\overline{p_i} = \begin{cases} H^l & if\, B_i^M = 0 \\ H^h & if\, B_i^M = 1 \end{cases} \tag{5}$$

where $\overline{p_i}$ is the decoded intensity of i^{th} pixel of the block. For even numbered pixels, some pre-defined interpolation equations are applied using the reconstructed odd pixel values. Thus, the complete hiding block is estimated as shown in block (f) of Fig. 1. In case of hiding blocks as well, the 4th pixel LSB is also marked to discriminate them from the saving block as shown in block (g). Now, the difference between the original and new blocks is calculated as shown in Fig. 1h and m for the hiding and saving block, respectively. Following that, the Huffman encoding is applied to the differences of all blocks to condense their size, resulting in the Huffman codes and dictionary as shown in Fig. 1n, which are then coupled with the encrypted secret data as depicted in Fig. 1o. On the other end, IAMBTC trios are stored into first three pixels. Afterwards, combined blocks are encrypted using a stream cypher with the assistance of a pseudorandom generator and an encryption key, except first MSB of 4th pixel of each block, as shown in block (i) and block (p) of Fig. 1. Finally, the remaining space of hiding blocks is used for embedding the Huffman code, directory and secret data, and an encrypted image with hidden secret data is obtained which is also known as marked encrypted image.

For image recovery and data extraction, the reverse of embedding process is performed, by the receivers. However, there may be three types of receivers. One who has only a data hiding key, he/she can only extract the Huffman code, Huffman directory and secret data and in case the receiver merely has an encryption key, he/she can distinguish between the hiding and saving blocks. Only the first three pixels are decrypted from the hiding block, which contains the trio $\{H^l,\ H^h,\ B^M\}$. The trio then can be used to predict all the pixels with the help of interpolation formulas defined in [22]. Saving blocks, on the other hand, are directly decrypted and an image is obtained. In case a receiver has both the keys, then he/she can extract the hidden information and can also restore the image using the same process defined above.

Both the works i.e., [22] and [23] are completely lossless and provides decent embedding capacity. Basically, most of the existing works which includes AM2PHC [23] and IAMBTC [22] either use linear prediction strategy or use block compression to exploit spatial correlation and create large room inside the image for embedding the secret information. However, none of the existing works seems to make use of variable size blocks/groups or chains based on the prevalent spatial correlation to make its optimal benefit. So, in this paper, we propose a new reversible data hiding method for encrypted images based on link chain strategy. The proposed method generates different size and shape link chains based on the prevalent spatial correlation and helps the data hider to embed large amount of secret data. In the next section, the proposed LCD-RDHEI method is described in detail.

3 Proposed method

This section presents the proposed method which can be seen with the purviews of content owner, data hider and recipient as follows. In the purview of content owner (3.1), operations such as image pre-processing, encryption of image, and process of supporting information (SI) embedding are done: and in the purview of data hider (3.2), operations such as extraction of PI, extraction of directory of link chain and embedding of the secret data are implemented; and in the purview of the recipient (3.3), operations such as extraction of the data and image decoding either separately or combined subject to the availability of hiding keys are performed. For clear representation of different purviews and operations there-in, Fig. 2 is provided that illustrates the framework of the proposed LCD-RDHEI.

As shown in Fig. 2, the content owner's major work is to encode the original image (I) by first pre-processing the image to create a room inside, for embedding the secret data. In the pre-processing, supporting information (SI) is generated which is embedded inside the image so that subsequent user i.e., data hider can extract to find the created room for embedding. Afterwards, the image is encrypted using an image key (K_e) to get an encrypted image [[I]] which is forwarded by the content owner to data hider, who may be present anywhere in the network. Now, the data hider identifies the pixels from [[I]] for embedding secret information that can be encrypted using a data key (K_d). Thus, a marked encrypted image $\overline{\overline{[[I]]}}$ is obtained which can be forwarded to receivers, who can extract the secret message and recover the image $\overline{\overline{(I)}}$ according to the availability of the keys i.e., K_d and K_e. In the next sub-section, the working process in the content owner's horizon is described in detail.

3.1 In content owner's Horizon

The content owner has three functions which include image pre-processing, encryption of the image and process of supporting information embedding, which are defined as follows:

3.1.1 Image pre-processing

The original image I of size $M \times N$ is scanned in raster scan manner and pixels are marked by their sequence number (S) of scan instead of their coordinates. The sequence number of a pixel $(p_{i,j})$ will be $((i-1)*N+j)$, where i and j represents row and column number. For example, the pixel $(p_{3,2})$ of the 512×512 size image can be represented with its sequence number i.e., $((3-1)*512+2) = 1026$, as (p_{1026}). Next, the content owner processes I to form highly correlated groups of pixels. The groups are hereafter called link chains (l_c) as these are also formed externally by making a link between consecutive pixels of the group. The pixels of the link chain are also called nodes. The first node of the link chain, which is also known as link chain head (h_{lc}), will have the sequence number of the corresponding pixel along with direction to the subsequent node and rest of the nodes only contain direction to their next node. It is to be noted that the link chains can be of any size and shape but will have pixels of very similar intensity or high spatial correlation. Thus, by doing this, a room can be created in the image for embedding the secret information. Next, processing of the untouched pixels which have not been part of any of the aforementioned link chain is done so that the reserved room can further be extended. Thus, the image pre-processing is done in two phases i.e., a) link chain directory generation and b) untouched pixel identification and their prediction which are discussed as follows:

a) **Link chain directory generation:** This phase has three steps which includes i) link chain generation, ii) categorization of link chains, iii) node representation in link

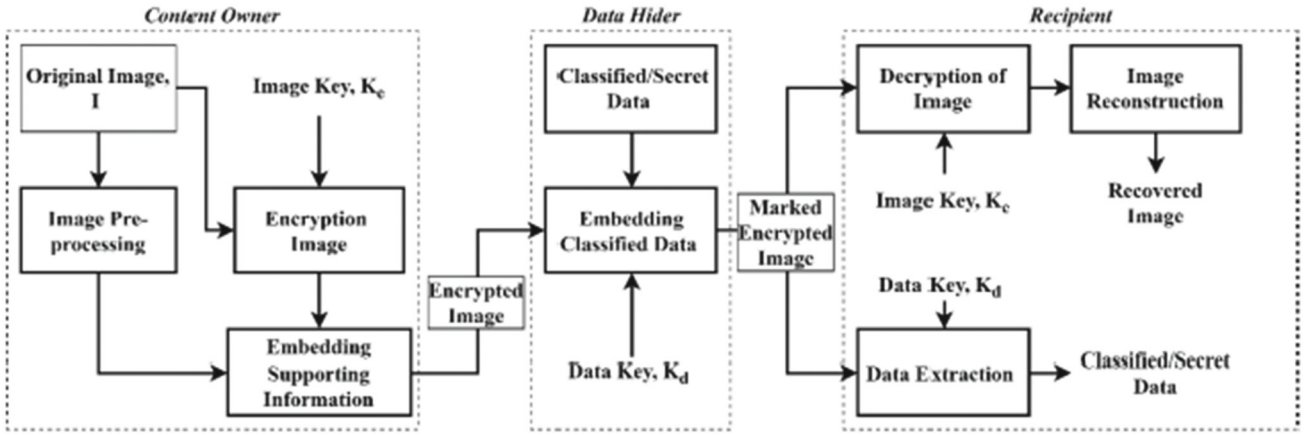


Fig. 2 Framework of the proposed LCD-RDHEI method

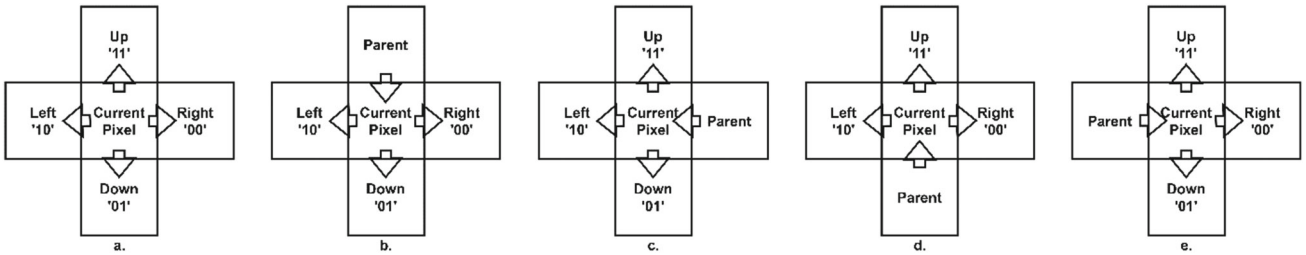


Fig. 3 An example of possible link movement for adjacent pixels (**a**) If the node is link chain head. **b** If the next node is down to the parent move, **c** If the next node is right to the parent move, **d** If the next node is up to the parent move, **e** if next node is left to the parent move

chains, which are introduced below. For more clarity, the algorithm to create the directory of link chains is provided in sub-section iv.

i) *Link chain generation:* For generating the link chains (l_c), first of all, a list (Li) is generated which stores all the pixels of I by their intensity value (val) and sequence number (S) in increasing order of their val (except the pixels reserved for process information (PI), which will be discussed later in Sect. 3.1.3). Now, the first pixel of the list (Li) is chosen as a link chain head (h_{lc}) and then any adjacent pixel from I that can be either right, down, left or up pixel, of the current pixel, is selected for adding the same into l_c, as per the Fig. 3a. The adjacent pixel can only join l_c, if the intensity difference of the h_{lc} and the adjacent pixel is less than a given user-defined threshold (i.e., variability coefficient (V_c)). The variability coefficient (V_c) shows the maximum allowable intensity difference within a l_c. The adding of the adjacent pixel to the current link chain may lead to a situation where multiple l_c are possible at any stage because more than one adjacent pixel i.e., right, down, left or up, can satisfy the variability coefficient condition. Thus, it may lead to a situation similar to the Hamiltonian path problem. Moreover, this problem is non-polynomial time solvable. So, it is not suitable for the content owner to find the maximum length l_c. To avoid the above situation, the proposed l_c generation method follows the greedy approach. The proposed method uses many recursive calls to determine the maximum length (max_len) for any l_c. In the recursive calls, multiple virtual l_c are formed at the step of adding an adjacent pixel, but only one virtual l_c with the maximum length is added as final, and the others are dropped. As a result, a pixel can be a member of many virtual l_c, requiring length calculations multiple times for a single pixel. To avoid the scenario, this method assumes that the first computation of the length of link chain from that pixel is perfect and that itself may be applied to any of the remaining virtual link chains from that pixel. Or, to put it another way, if virtual l_c generation has already been completed for any pixel, and any virtual l_c later reaches that pixel, that pixel is attached to the virtual l_c as the next node, and the already formed l_c is presumed to be the virtual l_c's tail. Although each virtual l_c must be preserved as part of the aforementioned process, as it can be used immediately during an interdicted call. As a result, this method uses two intermediate lists to store the intermediate information for each virtual l_c, where the first list (Len) contains the max_len of the l_c from the current pixel $I(S)$ and another lists (Dir) stores the encoded form of direction of the next possible

attached pixel that can be right (00), down (01), left (10) and up (11) obtained from the following Eq. (6).

$$S^c = \begin{cases} S+1 & if\, Dir(S) =' 00' \\ S+N & if\, Dir(S) =' 01' \\ S-1 & if\, Dir(S) =' 10' \\ S-N & if\, Dir(S) =' 11' \end{cases} \tag{7}$$

Function 1 $Max_Len_LC(I, C_s, X, V_c, S, Len^*, Dir^*)$

1.	***if*** $X \le I(S) < (X + V_c)$ $AND\ S \notin C_S$ $AND\ Len^*(S) = 0$	//returns true only if value of the adjacent pixel is within the defined range AND it does not belong to C_s AND the pixel is not already processed
2.	$Temp[1:4] = 0$	//initialize the Temp array of four elements to 0
3.	$Temp(1) = 1 + Max_Len_LC(I, \{C_S \cup S\}, X, V_c, S+1, Len^*, Dir^*)$	//computing max_length for right adjacent pixel
4.	$Temp(2) = 1 + Max_Len_LC(I, \{C_S \cup S\}, X, V_c, S+N, Len^*, Dir^*)$	//computing max_length for down adjacent pixel
5.	$Temp(3) = 1 + Max_Len_LC(I, \{C_S \cup S\}, X, V_c, S-1, Len^*, Dir^*)$	// computing max_length for left adjacent pixel
6.	$Temp(4) = 1 + Max_Len_LC(I, \{C_S \cup S\}, X, V_c, S-N, Len^*, Dir^*)$	// computing max_length for top adjacent pixel
7.	$[Len(S), ind] = MAX(Temp)$	// assigns max_length and index of the max_length adjacent pixel in $Len(S)$, *and* ind, respectively
8.	***if*** $ind = 1$ then $Dir^*(S) = '00'$; ***elseif*** $ind = 2$ then $Dir^*(S) = '01'$ ***elseif*** $ind = 3$ then $Dir^*(S) = '10'$ ***elseif*** $ind = 4$ then $Dir^*(S) = '11'$ ***endif***	//update the next pixel encoded direction in DIR using index of maximum value of Temp
9.	***endif***	
10.	***return*** $Len(S)$	// returns length

$$Dir(S) = \begin{cases} 00\ if\, S^c = S+1 \\ 01\ if\, S^c = S+N \\ 10\ if\, S^c = S-1 \\ 11\ if\, S^c = S-N \end{cases} \tag{6}$$

where S^c is the sequence of the next node of child node to be added in the l_c. Initially, Len and Dir are set null for every new l_c generation to avoid incorrect overlapping with the already generated link chains. Following the determination of a virtual l_c for a pixel, max_len and Dir of the next pixel is updated in Len and Dir, respectively, corresponding to the sequence number S. It can be observed from the *function-1:* $Max_Len_LC()$ that $Len(S)$ is checked from the Len before calculating the virtual l_c for a pixel $I(S)$ as all the intermediate virtual l_c share a common intermediate list. Additionally, pixels that are part of (any of) the link chains are regarded to be part of a cover set (C_S), which is initially set to null. The pseudocode to determine the maximum length of a link chain for a pixel $I(S)$ is given as function-1 (Max_Len_LC). It can be observed that maximum three recursive calls are possible in the above function Max_Len_LC for a pixel since one of the directions always points towards parent i.e., already a part of C_S, as illustrated in Fig. 3b–e. At last, the l_c is extracted from the intermediate lists as follows: first, a link chain head (h_{lc}) is constructed for the current pixel with sequence S, which also contains two bits of $Dir(S)$. The $Dir(S)$ gives the sequence number (S^c) of the next node (also called as child node) of the link chain as per the following Eq. (7).

Next, same process of adding the next node in the link list by considering S^c as the sequence number of the current pixel, is continued until the $Dir(S)$ is Null, which represents the last pixel of link chain. Finally, in the last node of the link chain, the encoded direction to the penultimate node is updated using Eq. (8) which is denoted by $S^p_{direction}$ whereas parent node is denoted as S^p.

$$S^p_{direction} = \begin{cases} 10\ if\, S^c = S+1 \\ 11\ if\, S^c = S+N \\ 00\ if\, S^c = S-1 \\ 01\ if\, S^c = S-N \end{cases} \tag{8}$$

Thus processing, all the link chains from the image can be extracted. In the next sub-section, categorization of the extracted link chains is discussed.

ii) *Categorization of Link chains:* The generated link chains are now categorized in three categories jump, small and large link chain, for the sake of achieving high embedding capacity and low distortion. Two user-defined thresholds i.e., T^1 and T^2 are used to separate them into the above suggested categories. More specifically, if the length (max_len) of any l_c is less than the first threshold (T^1) then it can be considered as a jump l_c, which can be represented by l_{cj} and if the length is greater than or equal to T^1 and less than T^2, it is considered small l_c which can be represented by l_{cs}, otherwise, large l_c which can be represented by l_{cl}. Note that, the T^1 and T^2 are fixed, i.e., $T^1 = 3$ and $T^2 = 16$ (for the sake of simplicity) in this paper. The proposed method does not mark the entries for jump link chains (l_{cj}) in

the cover set (C_s) so that those pixels can be utilized for other link chains (if possible). Therefore, there are only small (l_{cs}) and large (l_{cl}) link chains, effectively. All the generated link chains are sorted in increasing order based on their head node's sequence number.

iii) *Node representation in link chains:* For reserving bigger room, the pixel values of link chains are estimated either by their average values or by two quantization levels along with a bitmap based on the type of link chain. More specifically, the pixels of small link chains are estimated by their average (H) using Eq. 2, and pixels of large link chains are estimated by two quantization levels and bitmap using IAMBTC as defined in II-B. The process of IAMBTC has been adjusted for employing on link chains in place of image blocks. Firstly, H^l and H^h are calculated using Eqs. (2) and (3) for l_{cl}. Next, to represent every node of l_{cl} is generated using the process defined in II-B and every alternate (in other words, even-positioned) pixel representation in the bitmap is dropped, to get the partial bitmap $\left(B^M\right)$. Moreover, the proposed method stores the difference value $\left(D^{hl}\right)$ calculated as $D^{hl} = H^h - H^l$ in place of H^h to further reduce the required number of bits to represent the trio of H^l, H^h and B^M, for representing the nodes of the l_{cl}. Since the D^{hl} can range between 0 to $V_c - 1$, so fixed Huffman codes are derived to represent D^{hl}. However, it should be noticed that if $V_c = 1$, then every l_{cl} is treated like l_{cs}, because all of the pixels have the same intensity in the l_{cl} and their average will also be the same, so producing H^l, H^h and B^M will be pointless and waste a lot of space. The derived Huffman codes are provided in Table II which can be used by content owner, data hider and recipients. Further, binary representation of H for l_{cs} and H^l, D^{hl} & B^M for l_{cl}, are stored as a sequence of bits (H_{bs}) in the same order as the h_{lc}

b) Untouched pixel identification and their prediction: The pixels, which have not been covered in any of the link chains (or PI, which will be discussed later Sect. 3.1.3), are called untouched pixels $\left(ut_p\right)$. In the proposed work, untouched pixels are also being used for embedding the secret information in order to increase the embedding capacity of an image. For this, AM2PHC [23] (which has already been discussed in Sect. 2.1) have been employed. Prediction for remaining pixels and generation of Huffman code and rules, on the other hand, are done in the same way as in AM2PHC [23].

3.1.2 Encryption of the image

For encrypting the image contents, the content owner uses stream cipher to encrypt original image I. For this, initially, a matrix (m_e) of size M × N and list (l_e) of size $\lceil len(H_{bs})/8 \rceil$ are populated using a pseudo-random number generator considering image key (K_e) as the seed. Then, the image is encrypted using Eq. (9) and link chains are encrypted using Eq. (10), as follows.

$$[\![\mathrm{I}(i,\ j)]\!] = I(i,\ j) \oplus m_e(i,\ j) \quad \forall i,\ j \tag{9}$$

where $1 \le i \le M$, $1 \le j \le N$ and $[\![\mathrm{I}(i,\ j)]\!]$ is an encrypted pixel at coordinates $i,\ j$. After turning the l_e in a bit stream 8 bit for each integer, H_{bs} is encrypted as follows:

$$\overline{H}_{bs}(i) = H_{bs}(i) \oplus l_e(i) \tag{10}$$

where $\overline{H}_{bs}$ represents encrypted bit stream and $1 \le i \le size\ of\ H_{bs}$. Although, m_e and l_e can also be populated using a chaotic generator or AES algorithm in output feedback mode.

3.1.3 *Process of supporting information (SI) embedding*

Supporting Information (SI) is an additional information required by the receivers to extract the hidden secret data and decrypt the image. This information must be embedded in the image for the sake of hiding and extraction of secret data and decoding of the image. The details regarding supporting information and its embedding algorithm are presented in the following two subsections:

a) Details of supporting information (SI): *SI* basically, consists of following information components:

- **Variability coefficient (V_c):** The variability coefficient shows the maximum intensity difference (possible) within a l_c. Since the pixel values in a grayscale image, ranges from 0–255, the maximum possible absolute difference between any two pixels can be 255 which can be represented by 8 binary bits. Therefore, 8 bits are fixed for storing the value of V_c in the present work.
- **Location of the first link chain head $\left(d_l^1\right)$:** The location of the first link chain head which is denoted by d_l^1, is represented in $\lceil log_2(M \times N) \rceil$ bits.
- **Greatest location difference $\left(d_l^{max}\right)$and location difference sequence** (dl_{bs}): As mentioned in the preceding sub-Sect. 3.1.1, the link chains are sorted based on the location sequence number of their head nodes in a directory of link chains (DL_c). To get the greatest location difference between any two consecutive link chains, the location difference $\left(d_l^{k+1}\right)$ between every two consecutive heads i.e., h_{lc}^k and h_{lc}^{k+1}, is calculated using following Eq. (11).

$$d_l^{k+1} = \left(h_{lc}^{k+1}.S - h_{lc}^k.S\right) \tag{11}$$

where $h_{lc}^{k}.S$ and $h_{lc}^{k+1}.S$ represents the sequence number of head nodes of k^{th} and $(k+1)^{th}$ link chains. From the calculated difference sequences, the greatest location difference $\left(d_l^{max}\right)$ is identified and all the difference sequences are converted into $\lceil log_2\left(d_l^{max}\right)\rceil$ bit binary and stored in dl_{bs}.

- Huffman code rules (HCR): In the proposed work, multi-MSB matching [23] based data hiding scheme has been used for embedding in the untouched pixels $\left(ut_p\right)$ as described in II-B. In the scheme, a label map is generated which can have 9 different values. For optimally storing the label map, Huffman encoding has been used as described in [23]. For the Huffman encoding, some rules have been devised as in [23] which can be represented in 32 bits.
- **Encrypted link chain representation information $\left(\overline{H_{bs}}\right)$:** In Sect. 3.1.1, a bitstream of link chain representation information (H_{bs}) that contains average (H) for small link chains, low quantization $\left(H^l\right)$, quantization difference $\left(D^{hl}\right)$, and bit map $\left(B^M\right)$ for large link chains is generated. Afterwards, the bitstream H_{bs} is encrypted in Sect. 3.1.2 to get the encrypted bitstream of link chain representation information.
- **Encoded label map (ELM):** ELM is obtained form Sect. 3.1.1 for identification of untouched pixels and their prediction.
- **Encrypted MSBs of starting pixels ($EMSP$):** Six (6) MSBs of starting $\left(\left\lceil\frac{3\times log_2(M\times N)+11}{6}\right\rceil\right)$ pixels of image $[\![\mathrm{I}]\!]$ which are not part of any of the link chains are stored in $EMSP$. This information is required for recovering the image by the decoder and therefore needs to be recorded.

All of the above defined information components (if concatenated together) represents SI.

b) ***SI embedding algorithm:*** As discussed in the preceding subsection, the SI consists of several information components. However, these information components have been grouped in two different sets namely Process information (PI) and Auxiliary information (AI) based on their requirement and for the sake of simplicity to employ the proposed method. The AI is majorly required by the hider for embedding of the secret data and by the receivers as well during their process. In contrast, PI is required to initiate any of the phases such as hiding, decoding and/or extraction. More specifically, AI consists of concatenation of $dl_{bs}\left(\text{of size } number\ of\ n * \lceil log_2\left(d_l^{max}\right)\rceil \text{ bits}\right)$, HCR (of size 32 bits), ELM, $EMSP$ and $\overline{H}_{bs}$, in the given order. The size of ELM, $EMSP$ and $\overline{H}_{bs}$ are dependent on multiple parameters such as number of untouched pixels, PI, and number of link chains and their type etc. As far as PI is concerned, It is a concatenation of d_l^{max} (of size $\lceil log_2(M\times N)\rceil$ bits), d_l^1 (of size $\lceil log_2(M\times N)\rceil$ bits), V_c(of 8 bits), and LAI (o size $\lceil log_2(M\times N)\rceil$ bits) in the given order, where LAI represents length of AI. Thus, the total size of PI is $(3\times\lceil log_2(M\times N)\rceil+11)$ bits. The PI is basically embedded in the starting pixels of the encrypted image I in raster scan manner by replacing 6 MSBs of each pixel with the PI bitstream. These pixels can be called *PI enabled pixels* and their total count (NP_{PI}) is $\left(\left\lceil\frac{3\times log_2(M\times N)+11}{6}\right\rceil\right)$. Next, AI is embedded in the 6 MSBs of each pixels, which are part of link chains. The algorithm-2 to embed SI which includes both AI and PI, is provided as follows:

3.2 In the purview of data hider

Data hider is a person who basically embeds the secret information inside a cover media. Here, the cover media is an encrypted image with SI i.e., $\overline{[\![\mathrm{I}]\!]}$. Before starting the embedding of secret information, the data hider first needs to extract PI, link chains, and encoded label map (ELM), respectively as defined in the following subsections. Subsequently, he/she performs the embedding of secret data in the image using the process discussed in sub-Sect. 3.2.4.

3.2.1 Extraction of PI

For extraction of the PI, reading of 6 most significant bits (MSBs) in raster scan is started from the very first pixel of $\overline{[\![\mathrm{I}]\!]}$ and first of all, $\lceil log_2(M\times N)\rceil$ bits i.e., 19 bits in case 512×512 pixels image are extracted from the first four pixels to get greatest location difference $\left(d_l^{max}\right)$ by their decimal value. Then, next $\lceil log_2(M\times N)\rceil$ bits are extracted which are used to determine d_l^1 and subsequent 8 bits to get V_c. And lastly, next $\lceil log_2(M\times N)\rceil$ bits are used to determine the length of auxiliary information, which is represented by LAI. Thus, a total of $3\times\lceil log_2(M\times N)\rceil+11$ bits are read.

3.2.2 Extraction of link chains and embedding of secret data in link chain

The link chain extraction process is somewhat similar to the link chain generation process. Here, the data hider first reaches to the head of the first link chain using d_l^1, which is the pixel $[\![\bar{\mathrm{I}}(S)]\!]$ and reads 6 MSBs of the pixel (except the PI enabled ones) for EI (which is same as AI in the encoding/hiding phase, but being denoted differently for avoiding any confusion) extraction and two LSBs for direction to its next (or child) pixel of the link chain. Next, read again the 6MSBs of the next pixel and 2 LSBs for the direction to its child node using Eq. 7. This process is continued until complete auxiliary information is not retrieved. In case, the direction of the next (or child) points to its parent, which in turn signifies the end of the k^{th} link chain $\left(l_c^k\right)$, the next link chain $\left(l_c^{k+1}\right)$ needs to be picked. To get the next link chain head node location, the hider reads bits

$\beta+1\ to\ \beta+\lceil log_2(d_l^{max})\rceil$ and converts them into decimal to get d_l^{k+1}, where $\beta=(k-1)*\lceil log_2(d_l^{max})\rceil$. Next, d_l^{k+1} is added to the current link chain head node location to get the sequence number of $(k+1)^{\text{th}}$ link chain head. As soon as, all the bits of EI are extracted, the same process is repeated for embedding the bits of secret data by just substituting the process of reading the 6MSBs of the pixels by the replacing 6MSBs with the bits of secret data. The complete process is reiterated until we get $d_l^{k+1}=0$ which signifies that there is no more link chain in the image.

3.2.3 Embedding of secret data in remaining pixels

To embed the secret data in the untouched pixels, firstly dl_{bs}, HCR, ELM, $EMSP$ and $\overline{H}_{bs}$ are extracted from the EI. Next, the image is scanned in raster scan manner by skipping the PI enabled pixels and the pixels corresponding to the link chains for identifying the untouched pixels. Afterwards, information is embedded in each of the untouched pixels (ut_p) based on the corresponding label map information using AM2PHC [23] which is also reviewed in the Sect. 2.1.

Thus, all the pixels of the image are utilized for embedding the secret information. In the next subsection, the algorithm-3: for embedding of secret data is presented in step-wise manner.

3.3 In the purview of recipient

Since the proposed method is a separable RDHEI method, there can be two different receivers namely decoder and extractor. The decoder is the person who hold the image key (K_e), and can get back the contents of the original image whereas the extractor holds the extracting key (K_d) to get the hidden contents. It should be emphasised that the decoder and extractor can be the same person if they possess both of the aforementioned keys. In the following subsection, data extraction is defined first, followed by image decoding in the subsequent subsection.

3.3.1 Data extraction

The data extraction process is basically similar to the hiding process (described in Sect. 3.2). Here, the extractor first needs to extract the PI, link chain and AI like the hider and then reads the six bits from each pixel instead of replacing them as in the case of data hiding to get the encoded secret information from the link chain pixels. Next, the untouched pixels (ut_p) are processed using the label map to get back the hidden data. Finally, the extractor decodes the extracted information using the key (K_d) to get back the hidden information.

3.3.2 Image decoding

As discussed earlier, the image decoding process is performed by the decoder who holds the decoding key (K_e). In the image decoding process, first of all, PI, directory of link chains (DL_c), label map (LM) and AI is extracted on the similar lines of the hider (described in Sect. 3.2). Afterwards, dl_{bs}, HCR, ELM, $EMSP$ and $\overline{H}_{bs}$ are extracted as per the steps 7, 8 and 10 defined in the decoding algorithm. Next, the secret data is then extracted from the untouched pixels with the help of AI and PI. Afterwards, a pseudo-random number list $\tilde{l}_e$ is created using key using K_e with the help of a pseudo-random number generator and $\overline{H}_{bs}$ is decrypted using Eq. (12), as follows.

$$H_{bs}(i)=\overline{H}_{bs}(i)\oplus\tilde{l}_e(i) \tag{12}$$

where, $1\le i\le size\ of\ \overline{H}_{bs}$. Additionally, image decryption is done according to Eq. (13) by first creating a pseudo-random number matrix $\tilde{m}_e$ of size M × N size with the help of key (K_e) using a pseudo-random number generator.

$$\bar{\bar{\mathrm{I}}}(i,\ j)=\overline{\overline{[[\mathrm{I}]]}}(i,\ j)\oplus\tilde{m}_e(i,\ j)\forall i,\ j \tag{13}$$

where $1\le i\le M$ and $1\le j\le N$. Now to restore the image, all l_c of DL_c are processed in their order as follows: If the l_c is small then 8 bits are removed from H_{bs} and converted into decimal to get average (H). Afterwards, a list for recovered intensity (R) of size $max_len(l_c^k)$ is generated where each element of R is assigned average (H) value. In case the l_c is large then 8 bits are removed the from H_{bs} and converted into decimal which represents low quantization (H^l). Next, D^{hl} bits are removed from H_{bs} and decoded using Table II. Afterwards, high quantization (H^h) is computed as $H^h=H^l+D^{hl}$ and partial bitmap (B^M) is obtained by extracting $\max_len(l_c^k)/2$ bits from H_{bs}. The complete bitmap (B_c^M) is estimated using following Eq. (14)

$$B_c^M(x)=\begin{cases}B^M\left(\lceil\frac{x}{2}\rceil\right) & if\ x\%2=1,\\ -1 & if\ x\%2=0.\end{cases} \tag{14}$$

where x is a positive integer (1, 2, 3, ..., $\max_len(l_c)$). Thus, odd locations are filled by the values of the partial bitmap and '− 1' is set to the even locations of the bitmap. Next, a list (R) to recover intensity values of the pixels is constructed using Eq. (15) as follows:

$$R(x)=\begin{cases}H^l & if\ B_C^M(x)=0\\ H^h & if\ B_C^M(x)=1\\ R_avg & if\ B_C^M(x)=-1\ and\ x<\max_len(l_C)\\ R(x-1) & if\ B_C^M(x)=-1\ and\ x=\max_len(l_C)\end{cases} \tag{15}$$

where, $R_avg = \left\lfloor \frac{1}{2}(R(x-1) + R(x+1)) \right\rfloor$. All the pixels of $\bar{\bar{\mathrm{I}}}$ corresponding to l_c are then replaced by a corresponding value of recovered intensity (R). Thus, the pixels corresponding to the link chains are recovered. Next, the untouched (ut_p) pixels are reconstructed using AM2PHC [23] as described in II-A. Thus, the complete image is recovered and secret data is extracted.

4 Experimental results and discussion

This section elaborates upon experimental results and their analysis in order to compare and evaluate the performance of the proposed (LCD-RDHEI) method with respect to the existing and related RDHEI methods. The discussion and analysis of the experimental results has been done in four parts which are discussed in four sub-sections as follows. The first sub-section presents a detailed explanation of the LCD-RDHEI considering image 'Lena' as a use case, the second sub-section discusses performance analysis on different test images and datasets such as BOSSBase [35], BOWS-2 [36] and UCID [37], the third sub-section provides discussion around security of the method under consideration, and final sub-section presents a comparative analysis of the LCD-RDHEI with respect to the state of art methods. The main focus is on increasing the embedding capacity or embedding rate, illustrating how much data can be incorporated either in an entire image or per pixel. Additionally, PSNR, which is a logarithmic metric devised to measure the variations between an original image and its reconstructed version, is also explored. Infinite PSNR implies absolute reversibility, whereas superior to 30 dB is regarded as satisfactory reversibility.

4.1 A detailed explanation of the LCD-RDHEI considering image Lena as a use case

The trade-off of ER and PSNR i.e., if V_c increases the ER increases but PSNR decreases has been depicted in Table III, where ER and PSNR are dependent on the variability coefficient V_c. The proposed LCD-RDHEI method is applicable for $V_c = 1$ to 256, but for sake of simplicity and for comparative ease, $V_c = 1 - 18$ or PSNR $> 40dB$ has been taken into consideration in the presented experiments. However, the content owner can decide the value of V_c according to his/her application needs.

A test image 'Lena' is chosen to experimentally study different aspects such as pixel distribution of original and reconstructed image, representation of original, encrypted and reconstructed images etc., as provided in Fig. 4 for $V_c = 12$. In Fig. 4a, the original image of Lena of 512×512 size is shown and its pixel distribution using a histogram is

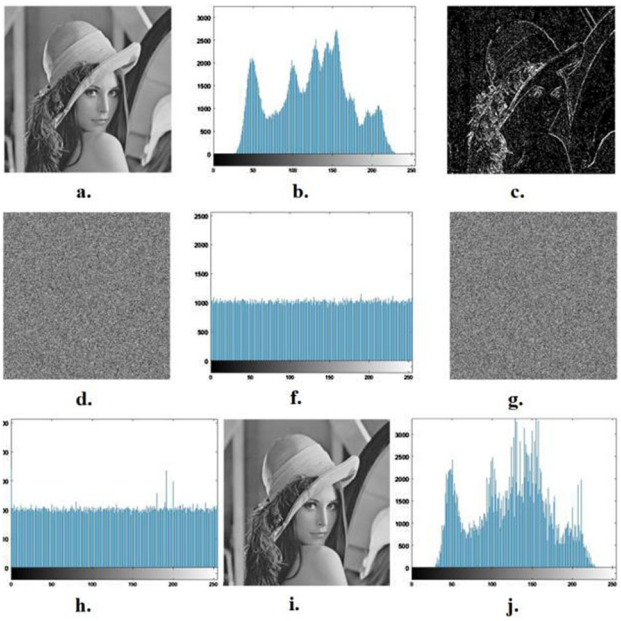


Fig. 4 Experimentation on image 'Lena' considering $V_c = 12$ (**a**) Test image; **b** Histogram of the test image (**c**) Coverage of Link Chains (**d**) Encrypted image (**e**) Histogram of encrypted image (**f**) Marked image (**g**) Histogram of Marked image (**h**) Reconstructed image (**i**) Histogram of the reconstructed image

presented in Fig. 4b. Since the proposed method starts its process with link chain (l_c) generation as discussed above, the coverage of pixels by link chains in the Lena image is also presented in Fig. 4c. The black colour in the Fig. 4c represents the pixel covered by the link chains. However, the LCD-RDHEI method generates a total of 13,585 l_c out of which 2666 are large l_c and remaining 10,919 are small l_c, where the maximum size of l_c thus formed is 1553 pixels whose initial location is at coordinate (233, 475) or sequence number (119,259). In addition to this, a total of 223,119 pixels have been covered in the link chain which includes 149,619 pixels of large l_c. Two bits are reserved for direction, therefore the preceding l_c can reserve six bits per pixel, or a total of 1,338,714 bits. Following that, other room in the image is reserved using AM2PHC on the remaining 39,025 pixels, and a label map of 158,814 bits is generated for a total room of 246,001 bits.

The total supporting information includes 85 bits of PI (that includes 18 bits for d_f^l size identification, 9 bits for d_f^l, 8 bits for V_c, 32 bits for Huffman code [23], 18 bits for (E^f)), 236,925 bits for all d_f^l, H, H^l and D^{hl} for all l_c, 74,807 bits for bitmap and 158,814 bits of label map, which will embed after encryption. In Fig. 4d, the encrypted image which has been encrypted by considering $K_d = 143$, is shown. The histogram of the encrypted image is presented in Fig. 4e, which clearly shows that pixels are equally distributed after encryption and there is no similarity left between the original

image i.e., Fig. 4a and encrypted image i.e., Fig. 4d. Then, 470,631 bits SI are embedded in total reserved room, which is the sum of the 1,338,714 bits generated by l_c's and the 246,001 bits generated by AM2PHC. Next, the data hider can embed secret data at a rate of 4.2498bpp in the aforesaid SI-enabled encrypted image. Figure 4f displays the marked encrypted image containing 1,114,084 bits of secret data. The pixel distribution of the marked encrypted image has been provided in Fig. 4g. This distribution is closer to the encrypted version provided in Fig. 4e which implies that the marked encrypted image also does not have any clue about the original image. Finally, the recovered image is shown in Fig. 4h after error-free extraction of the hidden data. The PSNR between the original image, shown in Fig. 4a, and the reconstructed image, shown in Fig. 4h, is 40.2966 dB. The histogram of the recovered image, which is substantially identical to the original image, is shown in Fig. 4i.

4.2 Performance analysis on different test images and datasets

In order to analyse the efficiency of the proposed LCD-RDHEI technique, the 8 standard test images namely Lena, Man, Baboon, Peppers, Airplane, Boat, Barbara and Crowd, each of size 512 $\times$ 512 pixels, have been taken into consideration. The experimental results in terms of ER (in bpp) and PSNR (in dB) by setting $V_c = 1$ to 16, have been presented in Table III. At $V_c = 1$, PSNR is infinite as the receiver can losslessly get the original image after extraction of the secret data. However, as the value of V_c increases, the PSNR decreases but embedding capacity increases and vice-versa is also true, which represents the traditional trade-off between the ER and PSNR. It can be observed from the Table III that the smother images such as Airplane, Lena provide more ER as compared to the rough images such as Baboon. In other words, the maximum ER i.e., 3.0789 is achieved in case of smoothest image i.e., Airplane and in contrast, minimum ER i.e., 1.0923 is provided by the most complex image i.e., Baboon with infinite PSNR. This suggests that the LCD-RDHEI method efficiently utilizes the spatial correlation. As a result, around half of the bits (i.e., 4 bits) in a smother image are easily utilized for embedding while keeping the PSNR over 40 dB. It can be observed that the in few cases, ER at $V_c = 2$ is lower than that of at $V_c = 1$ because LCD-RDHEI, at $V_c = 2$ and onwards, employs IAMBTC to represent the pixels of link chain by trio of H^l, D^{hl} and bitmap, which results in more overhead instead of one average value (in case of $V_c = 1$).

The performance of the LCD-RDHEI is also analysed on three data sets at $V_c = 1$ and 10. For this, the results of embedded capacity of each data set and corresponding PSNR are provided in Table 1. At $V_c = 10$, the ER for 500 randomly selected images from the BOSSBase [35], BOWS-2 [36], and UCID [37] are 4.2325, 4.0012, and 3.8497 respectively, while the PSNR values are 42.8769, 40.9913, and 41.0458. Conversely, at $V_c = 1$, the proposed method attains infinity PSNR for the similarly selected 500 images, delivering ER values of 3.2348, 3.0415, and 2.704 bpp for the BOSSBase, BOWS-2, and UCID databases respectively.

Table 1 Embedding Rate (bpp) and PSNR (dB) of the recovered images with $V_c = 1$ and $V_c = 10$ for three datasets

		BOSS Base	BOWS-2	UCID
V_c=1	ER	3.2348	3.0415	2.704
	PSNR	$+\infty$	$+\infty$	$+\infty$
V_c=10	ER	4.2325	4.0012	3.8497
	PSNR	42.8769	40.9913	41.0458

4.3 Comparison with State of Art methods

To comprehensively evaluate the performance of the proposed LCD-RDHEI method against some of the notable and recently developed existing methods such as the works of Fu et al. [32], Gao et al. [31], Liu et al.[30], Qin et al.[29], Li et al. [28], Gao et al. [27], Wang et al. [26], Xiao et al. [25], Yin et al. [23], Shiu et al. [22], Puteaux et al. [11], Cao et al. [9], Zang et al. [8] and Ma et al. [7], have been comparatively analysed. The aforementioned state-of-the art methods belong to both VRAE and RRBE. Figure 5 graphically represents the comparative analysis between the SOTA and the proposed method, focusing particularly on PSNR and ER as the pivotal performance metrics. The comparative analysis is given using four different characteristics test images Lena, Man, Baboon, and Airplane. As inferred from Fig. 5, the ability to embed a substantial payload is feasible in every case while maintaining an acceptable level of PSNR.

More specifically, the proposed method, in case of Lena and Man images which are moderately smooth images, surpasses all of the existing works in terms of the maximum ER but at the cost of PSNR. Further, the performance of the proposed method is at peak for highly smooth i.e., Airplane and highly complex image i.e., Baboon as evident from the Fig. 5c and d, respectively. For these images, the method seems to either beat all of the existing works or provide comparable results even at infinity PSNR. This has been made possible because of the method's effectiveness in exploiting the neighbourhood correlation. Overall.

Overall, the illustrative insights from Fig. 5 reveal that the Proposed LCD-RDHEI method achieves the highest ER, with negligible distortion. Moreover, its pivotal to highlight that, unlike the existing works which are usually constrained to a fixed embedding capacity, the proposed method offers enhanced flexibility. This adaptability allows for increased

capacity through manageable increments in distortion, providing a versatile solution that can be tuned for optimal performance.

5 Conclusion

In this research article, a high embedding capacity link chain driven RDHEI method based on pixel correlations, was proposed. The proposed LCD-RDHEI method collects the highly correlated pixels in terms of link chains which are of variable sizes and shapes. This collection of correlated pixels based on a variability co-efficient, helps in creating a big room before encryption which in turn helps in embedding the large amount of secret data. It is to be noted that the proposed method works as completely reversible data hiding method in encrypted domain if the variability co-efficient is zero, in other words, when all the pixels have the same intensity value within a link chain, which results into minimal embedding capacity. At larger embedding capacities where the values in a link chain are subject to a user-defined range based on the variability co-efficient, some of the contents of the cover media might be lost at the time data extraction and image restoration. To our knowledge, this is the first method that suggests the usage of variable size and shape of groups so that highly correlated groups can be formed. Experimentally, the proposed LCD-RDHEI method provides significantly greater embedding capacity than the existing and related data hiding methods. In future work, the focus

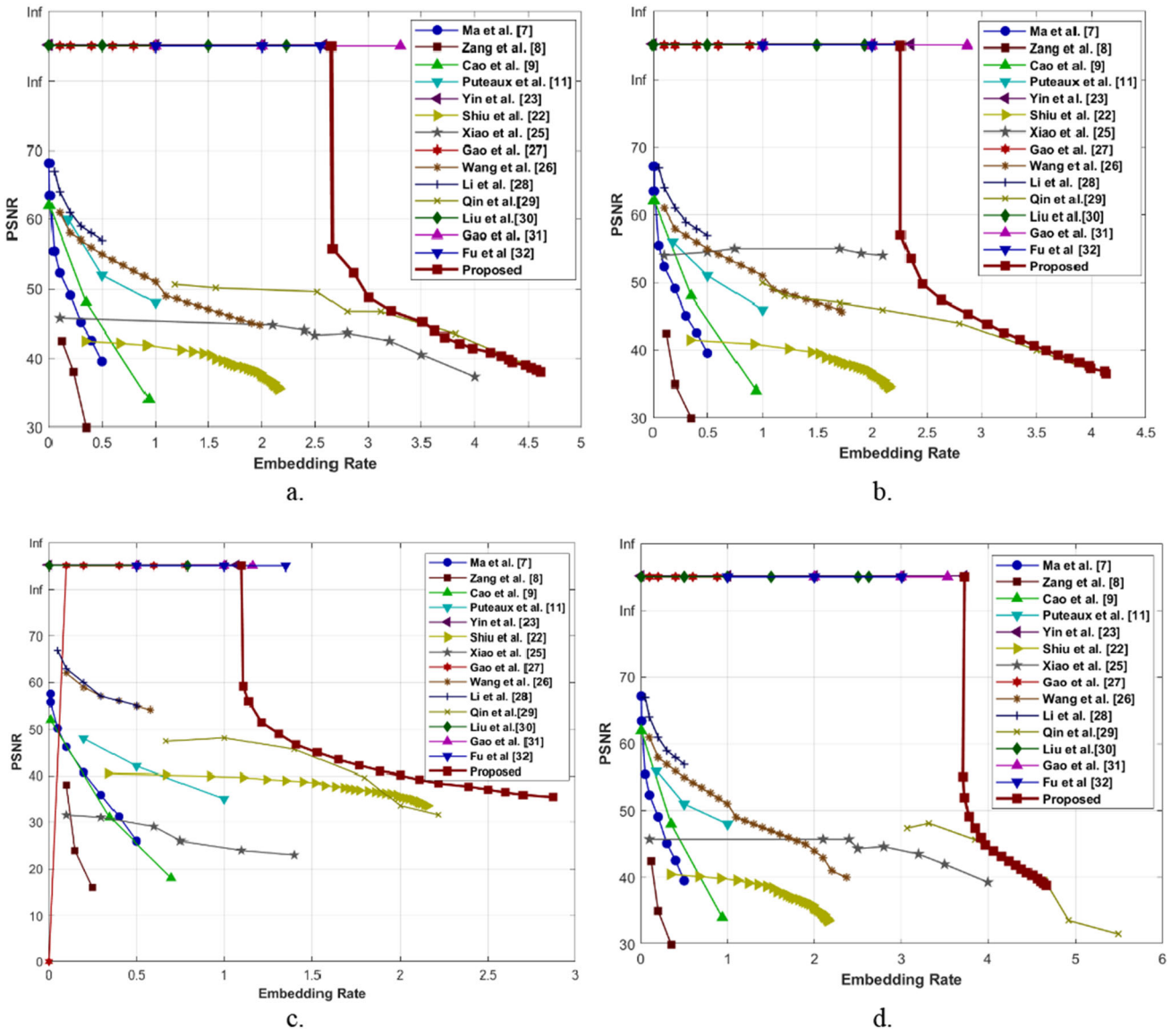


Fig. 5 Comparative analysis of proposed method w.r.t. SOTA methods for four test images of: **a**. Lena, **b**. Man, **c**. Baboon and **d**. Airplane

will be on reducing the distortion in case of long chains. Further, efforts will be made to reduce the overhead generated for identifying the next link chain.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11760-024-03275-1.

Author contributions Ankur and R.K. wrote the main manuscript text and conceptualized the idea. AKS reviewed the manuscript and provided supervision.

Data availability The data can made available on request to the first author.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Tchernykh, A., Schwiegelsohn, U., Talbi, E.G., Babenko, M.: Towards understanding uncertainty in cloud computing with risks of confidentiality, integrity, and availability. J. Comput. Sci. **36**, 100581 (2019)
2. Ali, M., Khan, S.U., Vasilakos, A.V.: Security in cloud computing: Opportunities and challenges. Inf. Sci. **305**, 357–383 (2015)
3. Bender, W., Gruhl, D., Morimoto, N., Lu, A.: Techniques for data hiding. IBM Syst. J. **35**(3.4), 313–336 (1996)
4. Puteaux, P., Ong, S., Wong, K., Puech, W.: A survey of reversible data hiding in encrypted images–The first 12 years. J. Vis. Commun. Image Represent. **77**, 103085 (2021)
5. Kumar, R., Chand, S.: A reversible high capacity data hiding scheme using pixel value adjusting feature. Multimed. Tools Appl. **75**, 241–259 (2016)
6. Ni, Z., Shi, Y.Q., Ansari, N., Su, W.: Reversible data hiding. IEEE Trans. Circuits Syst. Video Technol. **16**(3), 354–362 (2006)
7. Puech, W., Chaumont, M., Strauss, O.: A reversible data hiding method for encrypted images. In: Security, forensics, steganography, and watermarking of multimedia contents X, Vol. 6819, pp. 534–542. SPIE (2008, March)
8. Zhang, X.: Reversible data hiding in encrypted image. IEEE Signal Process. Lett. **18**(4), 255–258 (2011)
9. Hong, W., Chen, T.S., Wu, H.Y.: An improved reversible data hiding in encrypted images using side match. IEEE Signal Process. Lett. **19**(4), 199–202 (2012)
10. Wu, X., Sun, W.: High-capacity reversible data hiding in encrypted images by prediction error. Signal Process. **104**, 387–400 (2014)
11. Ankur, Kumar, R., Sharma, A. K.: Adaptive Two-Stage Reversible Data Hiding in Encrypted Images using Prediction Error Expansion. In: 2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC), pp. 427–432. IEEE (2023, May)
12. Qian, Z., Zhang, X.: Reversible data hiding in encrypted images with distributed source encoding. IEEE Trans. Circuits Syst. Video Technol. **26**(4), 636–646 (2015)
13. Qin, C., Qian, X., Hong, W., Zhang, X.: An efficient coding scheme for reversible data hiding in encrypted image with redundancy transfer. Inf. Sci. **487**, 176–192 (2019)
14. Tsai, C.S., Zhang, Y.S., Weng, C.Y.: Separable reversible data hiding in encrypted images based on Paillier cryptosystem. Multimed. Tools Appl. **81**(13), 18807–18827 (2022)
15. Ma, K., Zhang, W., Zhao, X., Yu, N., Li, F.: Reversible data hiding in encrypted images by reserving room before encryption. IEEE Trans. Inf. Forensics Secur. **8**(3), 553–562 (2013)
16. Zhang, X., Long, J., Wang, Z., Cheng, H.: Lossless and reversible data hiding in encrypted images with public-key cryptography. IEEE Trans. Circuits Syst. Video Technol. **26**(9), 1622–1631 (2015)
17. Cao, X., Du, L., Wei, X., Meng, D., Guo, X.: High capacity reversible data hiding in encrypted images by patch-level sparse representation. IEEE Trans. Cybern. **46**(5), 1132–1143 (2015)
18. Yi, S., Zhou, Y.: Binary-block embedding for reversible data hiding in encrypted images. Signal Process. **133**, 40–51 (2017)
19. Puteaux, P., Puech, W.: An efficient MSB prediction-based method for high-capacity reversible data hiding in encrypted images. IEEE Trans. Inf. Forensics Secur. **13**(7), 1670–1681 (2018)
20. Puyang, Y., Yin, Z., Qian, Z.: Reversible data hiding in encrypted images with two-MSB prediction. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7. IEEE (2018, December)
21. Chen, K., Chang, C.C.: High-capacity reversible data hiding in encrypted images based on extended run-length coding and block-based MSB plane rearrangement. J. Vis. Commun. Image Represent. **58**, 334–344 (2019)
22. Shiu, P.F., Tai, W.L., Jan, J.K., Chang, C.C., Lin, C.C.: An interpolative AMBTC-based high-payload RDH scheme for encrypted images. Signal Processing: Image Commun. **74**, 64–77 (2019)
23. Yin, Z., Xiang, Y., Zhang, X.: Reversible data hiding in encrypted images based on multi-MSB prediction and Huffman coding. IEEE Trans. Multimedia **22**(4), 874–884 (2019)
24. Wu, Y., Xiang, Y., Guo, Y., Tang, J., Yin, Z.: An improved reversible data hiding in encrypted images using parametric binary tree labeling. IEEE Trans. Multimed. **22**(8), 1929–1938 (2019)
25. Xiao, D., Li, F., Wang, M., Zheng, H.: A novel high-capacity data hiding in encrypted images based on compressive sensing progressive recovery. IEEE Signal Process. Lett. **27**, 296–300 (2020)
26. Weng, S., Zhang, C., Zhang, T., Chen, K.: High capacity reversible data hiding in encrypted images using SIBRW and GCC. J. Vis. Commun. Image Represent. **75**, 102932 (2021)
27. Gao, G., Tong, S., Xia, Z., Shi, Y.: A Universal Reversible Data Hiding Method in Encrypted Image Based on MSB Prediction and Error Embedding. IEEE Trans. Cloud Comput. **11**(2), 1692–1706 (2022)
28. Li, F., Zhu, H., Qin, C.: Reversible data hiding in encrypted images using median prediction and bit plane cycling-XOR. Multimed. Tools Appl. **82**(4), 6013–6032 (2023)
29. Qin, B., Zheng, H., Xiao, D.: High-capacity data hiding in encrypted images based on compressive sensing and intra-block difference coding. Applied Intelligence, 1–15, (2023)
30. Liu, M., Wang, K., Gao, T.: High-capacity reversible data hiding in encrypted images based on adaptive arithmetic coding and static Huffman coding. Cluster Computing, 1–19, (2022)
31. Gao, H., Zhang, X., Gao, T.: Hierarchical reversible data hiding in encrypted images based on multiple linear regressions and multiple bits prediction. Multimed. Tools Appl. **83**(3), 8757–8783 (2024)
32. Fu, Z., Chai, X., Tang, Z., He, X., Gan, Z., Cao, G.: Adaptive embedding combining LBE and IBBE for high-capacity reversible data hiding in encrypted images. Signal Process. **216**, 109299 (2024)
33. Kaspar Raj, I.: Image data hiding in images based on interpolative absolute moment block truncation coding. Math. Modelling Sci. Comput. **283**, 456–463 (2012)

34. Lema, M., Mitchell, O.: Absolute moment block truncation coding and its application to color images. IEEE Trans. Commun. **32**(10), 1148–1157 (1984)
35. Bas, P., Filler, T., Pevný, T.: Break our steganographic system": the ins and outs of organizing BOSS. In: Filler, T., Pevný, T., Craver, S., Ker, A. (eds.) International workshop on information hiding, pp. 59–70. Springer, Berlin (2011)
36. Bas, P., Furon, T.: Image database of BOWS-2. Accessed: Jun, 20, 2016–2017 (2017)
37. Schaefer, G., Stich, M.: UCID: An uncompressed color image database. In: Storage and Retrieval Methods and Applications for Multimedia 2004, Vol. 5307, pp. 472–480. SPIE. (2003)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Biotechnol Lett
https://doi.org/10.1007/s10529-024-03499-8

REVIEW

Machine learning: an advancement in biochemical engineering

Ritika Saha · Ashutosh Chauhan · Smita Rastogi Verma

Received: 17 November 2023 / Revised: 24 February 2024 / Accepted: 18 May 2024
© The Author(s), under exclusive licence to Springer Nature B.V. 2024

Abstract One of the most remarkable techniques recently introduced into the field of bioprocess engineering is machine learning. Bioprocess engineering has drawn much attention due to its vast application in different domains like biopharmaceuticals, fossil fuel alternatives, environmental remediation, and food and beverage industry, etc. However, due to their unpredictable mechanisms, they are very often challenging to optimize. Furthermore, biological systems are extremely complicated; hence, machine learning algorithms could potentially be utilized to improve and build new biotechnological processes. Gaining insight into the fundamental mathematical understanding of commonly used machine learning algorithms, including Support Vector Machine, Principal Component Analysis, Partial Least Squares and Reinforcement Learning, the present study aims to discuss various case studies related to the application of machine learning in bioprocess engineering. Recent advancements as well as challenges posed in this area along with their potential solutions are also presented.

R. Saha · A. Chauhan · S. Rastogi Verma (✉)
Department of Biotechnology, Delhi Technological University, New Delhi 110042, India
e-mail: smitar@dtu.ac.in

R. Saha
e-mail: ritikasaha_bt20a16_19@dtu.ac.in

A. Chauhan
e-mail: ashutoshchauhan_bt20a16_29@dtu.ac.in

Keywords Deep learning · Machine learning · Partial least squares · Principal component analysis · Reinforcement learning · Support vector machine

Mathematical notations and equations

(1) $y_{i\in N} = \left(y_{i^1}, \ldots, y_{ij}\right)^T$,
(2) $X_{i\in N} = \left(x_{i1}, \ldots, x_{iK}\right)^T$,
(3) $x_{ik} = a_{k_1} y_{i_1} + \cdots + a_{k_y} y_{i_J}$,
(4) $\sum_{i=1}^{N} \left(y_i k\right)^2 = y_{1k}^2 + \cdots + y_{ik}^2$,
(5) $a_{k1}^2 + \cdots + a_{kJ}^2 = 1$,
(6) $a_k a_m^T = 0, m = 1,2, \ldots, k-1$,
(7) y_{ij} represents the original variable with y_j in the ith data point, and x_{ik} refers to the latent variable x_k in the ith data point,
(8) $a_k = (a_{k1}, a_{k2}, \ldots, a_{kJ})$, y_k,
(9) X_i, R^J and Y_i, R^K ($K \in J$),
(10) Z_i, R^p and Q_i, R^l $p \in J$ and $l \in K$, Z_i and Q_i,
(11) Transition function T(s, a, s′) for all $s \in S$ and $a \in A$, a reward function R(s, a), and a discount factor γ, policy π(als).

Introduction

Bioprocess engineering is a multidisciplinary field with applications ranging from the synthesis of medicinal proteins, polysaccharides, vaccines, and antibiotics to products that can regulate harmful

Published online: 21 June 2024

wastes (Mowbray et al. 2021a, b). Current research focuses on developing mathematical models in subfields of this discipline (Esfandyari et al. 2017). These unique and dynamic mathematical concepts are used to optimize different kinetic models and further compute fluid dynamics for bioreactor and fermentation scale-up. Although a high amount of research has been done related to the development of complex mathematical theories, the research community still lacks efficient and robust modelling methods to accurately compute advanced biochemical engineering domain knowledge in terms of mathematical concepts (Ab Azar et al. 2020). Complex biochemical processes are often modelled with the help of physical, chemical, and biological principles, although we are slowly moving towards hybrid modelling, which encompasses statistical, machine learning (ML), and data-driven approaches to solve complex biochemical processes (Bhange et al. 2023; Cutajar et al. 2019; Vega-Ramon et al. 2021). With the help of data-driven and ML techniques, complicated correlations in the data can be uncovered (Khanal et al. 2023).

Biochemical processes, "one of the most complicated processes," have attracted scientists from different domains to work and contribute towards this research community (Dürrenmatt and Gujer 2011). There are different factors which determine the ambiguity of biochemical processes, such as biological cell materials, temperature, pH, control algorithms, sensor development, and the multiphasic nature of bioreactors, etc. These factors are often considered as a form of input for ML models, including Support Vector Machine (SVM), Artificial Neural Network (ANN), Multivariate Statistical Analysis (MSA), and Reinforcement Learning (RL), etc. ML can be used to optimize bioprocesses by modelling nonlinear relationships between operating parameters, for example, in biofuel production, optimization of inoculation proportion, bioreactor optimization, etc. (Kell 2006).

In the recent advancements of ML in biochemical and bioprocess engineering, diverse research studies have showcased the transformative impact of computational approaches on optimizing complex biological systems (Lee et al. 2018). The fermentation of blueberry juice, as explored in one study, utilized a unique combination of strains and intricate mathematical tools such as sequential simplex lattice mixture design, regression modelling, and genetic algorithm optimization (Liao et al. 2023). The ML-driven approach for predicting glucose responses from Kraft waste-based pretreatments exemplifies the seamless integration of artificial intelligence (AI) into bioprocess optimization, with high R^2 scores and robust validation metrics indicating the efficacy of the developed models (David et al. 2023). The exploration of pharmaceutical bioprocesses within the quality by design (QbD) framework employs the statistical design of experiments (DoE) as a robust tool, emphasizing the advantages of leveraging DoE and design space methodologies (Kasemiire et al. 2021; Lutnick et al. 2023). The study not only underscores the necessity for a reliable and efficient methodology to monitor and control various factors involved in optimizing bioprocesses but also demonstrates the shifting paradigm from traditional quality-by-testing (QbT) approaches to the more efficient QbD methodologies facilitated by ML techniques (Helmy et al. 2023). Furthermore, the study on improving Computational Fluid Dynamics (CFD) modelling in cell culture bioreactors introduces an ANN-based ML model (Cantarero-Rivera et al. 2024). This novel approach effectively corrects coarse-mesh-induced errors in CFD modelling, showcasing the potential of ML to enhance computational efficiency in understanding fluid dynamics within bioprocess systems. The study on monoclonal antibody (mAb) manufacturing in the pharmaceutical field explores the early stages of ML support for QbD-based development (Khuat et al. 2024). ML-based models, replacing conventional multivariate data analysis methods, are highlighted for their potential in contributing to reducing production times without compromising product quality, a critical aspect in the biopharmaceutical industry (Ge et al. 2017; Xu et al. 2014). Additionally, the study on biohydrogen production using dark fermentation introduces ML algorithms such as SVM, random forest (RF), ANN, and genetic algorithm (GA) to optimize and control this complex biological process (Nagata and Chu 2003; Pandey et al. 2023). The application of ML in predicting the growth kinetics of acetic acid-producing bacteria isolated from fruit waste further underscores the versatility of these techniques in various bioprocess applications (Upadhyay et al. 2024).

The present study aims to provide a fundamental mathematical understanding of commonly used ML and deep learning algorithms in the field of biochemical engineering, such as MSA, SVM, ANN, and RL.

Here, we present various case studies associated with the use of ML in bioprocess engineering. The impact of ML and the difficulties encountered are discussed in detail.

Machine learning techniques

ML algorithms employed in the field of biochemical engineering predominantly include MSA, SVM, ANN, and RL. To apply them to the field of biochemical engineering, a basic mathematical understanding of these algorithms is imperative.

Multivariate statistical analysis (MSA) or multivariate analysis (MVA)

"Multivariate Statistical Analysis (MSA)," which emerged in the 1990s, is a methodical approach for examining data comprising diverse measurements or observations. It involves the simultaneous examination of multiple dependent variables alongside other relevant factors. In the realm of mathematical theory, MSA relies on intricate computations to arrive at accurate conclusions (van Heel 1991). It is a versatile tool for handling situations where a multitude of factors impact the variability of dependent variables, contributing to more realistic and accurate outcomes. Examples of specific techniques within MSA include Multiple Regression Analysis, Conjoint Analysis, and Multiple Discriminant Analysis. Multiple Regression extends the principles of simple linear regression to predict a variable based on multiple predictor variables. Conjoint Analysis, often employed in market research, utilizes surveys to gauge the relative importance of various attributes of a product or service to consumers. Multiple Discriminant Analysis aids in classifying samples into groups based on predictor variables, with the goal of creating a model that minimizes misclassification.

MSA plays a pivotal role in supporting various crucial activities in bioprocessing, such as process characterization, scale-up, monitoring, fault diagnosis, and root cause analysis (Li et al. 2020). Rathore and Singh (2015) highlighted the utility of MSA as a diagnostic tool for identifying scale-up discrepancies and parameter interactions that adversely affect cell culture process performance. Through multivariate data analysis and modeling utilizing data from "small-scale (2 L), pilot-scale (2,000 L), and commercial-scale (15,000 L) batches", the root cause was successfully pinpointed, and experimental conditions were tailored to showcase and rectify it. The parameters under scrutiny in the investigation encompassed bioreactor p_{CO_2}, glucose, ammonium, lactate, raw materials, and seed inoculum, while the outcome parameters included product titer, attributes, cell viability, viable osmolality, and density.

MSA uses a broad variety of statistical techniques, including principal component analysis (PCA), partial least squares (PLS), and polynomial regression. Traditional MSAs are often regarded as "old" statistical procedures but have a wide range of uses in the bioprocessing sector. PCA and PLS help to minimize the number of dimensions in a high-dimensional dataset.

Principal component analysis (PCA)

PCA emerges as a critical tool to address the challenges posed by the "curse of dimensionality," particularly in the realm of high-dimensional datasets. The curse of dimensionality is characterized by issues such as overfitting, prolonged computation time, and diminished accuracy in ML models as the number of features or dimensions increases exponentially. Due to the expansion of dimensionality, the computational complexity and cost escalate, making tasks like clustering and classification challenging (Lever et al. 2017). Feature engineering techniques, including feature selection and feature extraction, aim to mitigate these challenges. PCA tackles dimensionality reduction by transforming a set of correlated variables into a set of uncorrelated variables.

PCA operates on the premise that while mapping data from a higher-dimensional space to a lower-dimensional space, the variance of the data in the latter should be maximized. This statistical procedure involves several key steps, starting with the standardization of the dataset to ensure a mean of 0 and a standard deviation of 1. Subsequently, the covariance matrix is computed to measure the joint variability between variables. The eigenvalues and eigenvectors of the covariance matrix are then calculated to identify the principal components. The essence of PCA lies in its linear transformation, which preserves the most variance in the data with the least number of

dimensions. This involves projecting the data onto selected principal components, thereby achieving dimensionality reduction. Practical implementation entails steps such as standardizing the dataset, computing the covariance matrix, determining eigenvalues and eigenvectors, sorting them in descending order, calculating explained variance, and selecting the number of principal components based on desired variance retention. Ultimately, the data is mapped onto the principal components, effectively reducing its dimensionality.

In PCA, the n-dimensional dataset is converted to m-dimensional feature space, where the dimensions are orthogonal. This algorithm works on a concept called the "Eigenvalue Decomposition" of a covariance matrix, which is calculated using a specific dataset. Figure 1 shows the steps involved in the PCA algorithm.

According to PCA, for each dataset, $y_{i\in N} = \left(y_{i^1}, \ldots y_{ij}\right)^T$ there are y different variables. The objective of this algorithm is to formulate $K<J$ new variables $X_{i\in N} = \left(x_{i1}, \ldots, x_{iK}\right)^T$ that satisfy the following conditions:

$$x_{ik} = a_{k_1} y_{i_1} + \cdots + a_{k_y} y_{i_j}, \tag{1}$$

where y_{ij} represents the original variable with y_j in the ith data point, and x_{ik} refers to the latent variable x_k in the ith data point.

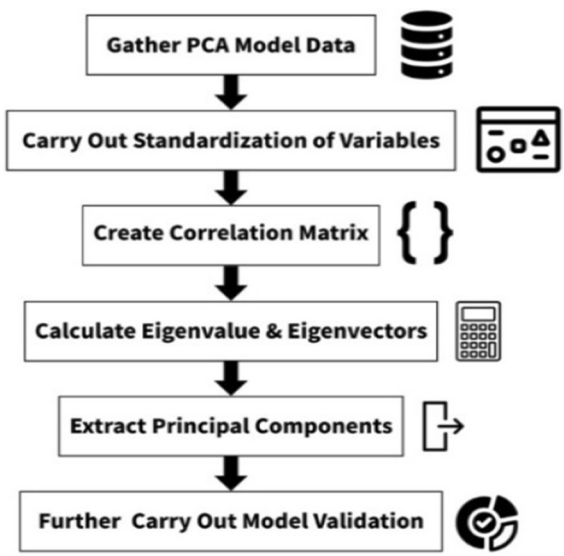


Fig. 1 Working of principal component analysis (PCA)

For the application of PCA in the field of biochemical engineering, the original variable y_j represents a particular variable with significance in a biochemical process. It may represent a bioprocess factor in the bioreactor dataset or an estimation at a specific wavelength in the spectroscopic dataset. These engineering parameters can include pH, temperature, flow rate, fluid velocity, shear stress, and mass transfer, etc.

The vector $a_k = (a_{kl}, a_{k2},....,a_{kJ})$ is obtained by increasing the correlation variable variance (y_k), as given in the following equations:

$$\sum_{i=1}^{N} \left(y_i k\right)^2 = y_{1k}^2 + \cdots + y_{ik}^2, \tag{2}$$

$$a_{k1}^2 + \cdots + a_{kJ}^2 = 1, \tag{3}$$

$$a_k a_m^T = 0, \quad m = 1,2, \ldots, k-1. \tag{4}$$

The purpose of PCA is to generate new variables by using different combinations of original dataset values, and this newly generated data is called the principal component. In most cases, maximum variations are mapped by PC1 (principal component 1) and PC2 (principal component 2). Hence, the data should be segregated into their PC sets (1st PC, 2nd PC) and mapped out to find possible connections (Abdi and Williams 2010).

PCA has been used in Raman Spectroscopy data analysis to uncover hidden relationships between various reactor operating variables in the case of biogas and biofuel production and is also used to identify process irregularities in large-scale cell culture systems (Goodacre 2003; Sharma et al. 2023).

Partial least squares (PLS)

PLS, also known as "projection to latent structure," emerges as a potent predictive modeling approach, particularly when conventional techniques like ordinary least squares (OLS) regression, canonical correlation, or structural equation modeling encounter difficulties arising from highly correlated predictor variables or an imbalance between predictors and cases (Boulesteix and Strimmer 2006). PLS methodology integrates components from PCA and multiple regression. In its operation, PLS begins by extracting a set of latent factors, aiming to capture the maximum

covariance between independent and dependent variables. This latent factor extraction proves pivotal in situations where traditional methods struggle due to multicollinearity or a surplus of predictors relative to observations. Subsequently, PLS employs a regression step to forecast the dependent variables' values using the decomposition of the independent variables (Wold et al. 2001).

Unlike traditional linear regression, PLS regression involves projecting both the predicted and observable variables into a new space. This unique characteristic categorizes the PLS family of methods as bilinear factor models. Particularly noteworthy is Partial Least Squares Discriminant Analysis (PLS-DA), a variant of PLS used when the dependent variable (Y) is categorical. The primary objective of PLS is to uncover fundamental relationships between two matrices, denoted as X and Y. It adopts a latent variable approach to model the covariance structures in these two spaces. In the X space, PLS endeavours to identify a multidimensional direction that explains the maximum variance in the corresponding multidimensional direction in the Y space. This approach makes PLS regression well-suited for situations where the matrix of predictors has more variables than observations and multicollinearity among X values is prevalent, scenarios in which standard regression may falter unless regularization techniques are applied.

If the original input and output dataset variables are X_i, R^J and Y_i, R^K ($K \in J$), then PLS transforms them into their respective latent variables Z_i, R^p and Q_i, R^l with $p \in J$ and $l \in K$. Furthermore, PLS obtains a correlation between Z_i and Q_i. Equations (2)–(4) are used to calculate Z_i and Q_i.

If there are several output variables to be correlated with input factors, it is usual practice to create multiple independent PLS models concurrently. Figure 2a and b depict the flow and purpose of the PLS algorithm, respectively. Figure 2c shows the working of PCA in association with the procedure of the PLS algorithm.

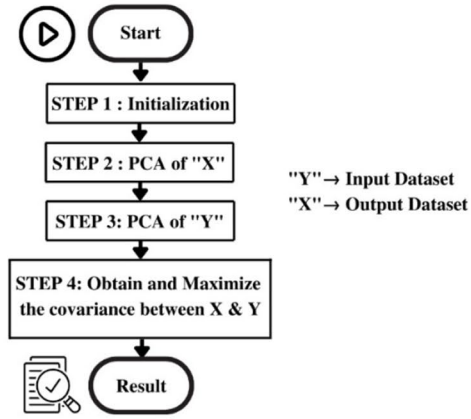
Since the PLS algorithm can handle missing data values, it has been used for analysing industrial batch data. Data is obtained from multiple batches and covariance is determined between input and output variables; this ability makes it suitable for bioprocess modelling (Patel et al. 2020). PLS is also employed for state estimation and control in bioreactors, as well as monitoring of bioprocesses (Mandenius and Titchener-Hooker 2013). It also serves several purposes in mapping protein–amino acid sequences and discovering connections between reactions and metabolic pathways (Whittle 1952).

Support vector machine (SVM)

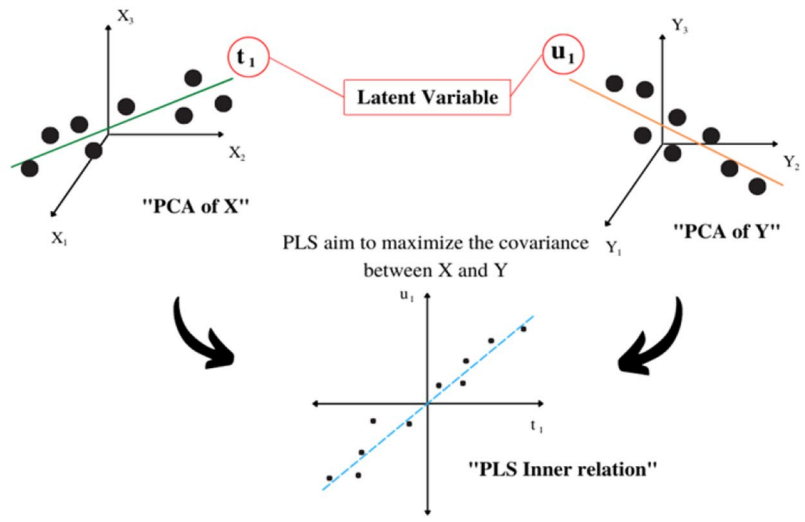
Statistical learning theory serves as the cornerstone for the innovative ML method known as SVM. It has become the ML community's research focus due to its higher accuracy. SVM is a powerful ML algorithm that operates on the principle of finding an optimal hyperplane to separate different classes in a feature space. Unlike many ML algorithms, SVM does not rely on explicit equations for class boundaries (Yang 2004). Instead, it focuses on finding the optimal hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points from each class. This is achieved using support vectors, which are the data points closest to the decision boundary. SVM's mathematical elegance lies in its ability to efficiently handle linear and non-linear relationships through the kernel trick without explicitly requiring equations for decision boundaries. The algorithm's strength lies in its geometric interpretation and the optimization of margins, providing a versatile tool for various classification tasks.

Figure 3 provides a clear visual representation of the SVM applied to two-dimensional data. In this illustration, two distinct lines are constructed to separate data points belonging to different classes. These lines, also known as decision boundaries, are strategically placed to maximize the margin between the two classes, ensuring the obtention of the best possible hyperplane. SVM's objective is to find this hyperplane, which not only separates the data but also maximizes the distance between the nearest data points from each class. This approach results in an optimal classification model, making SVM a powerful tool for various applications, including image recognition, text classification, medical diagnosis, and financial forecasting, etc.

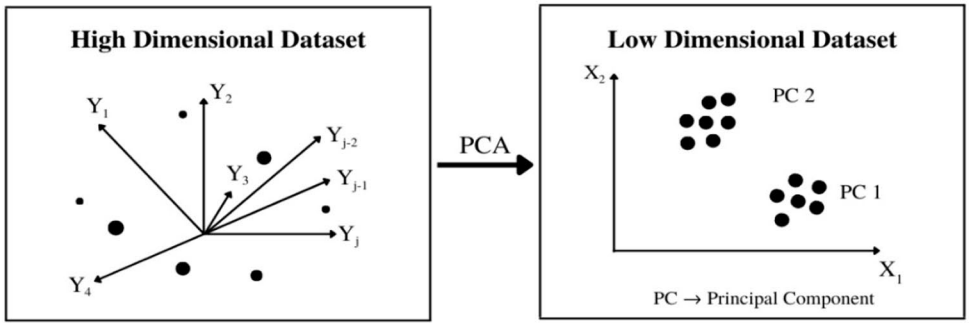
Since the late 1990s and early 2000s, SVMs have been used for both classification and regression problems in the field of biochemical engineering. SVMs extract intricate nonlinear correlations frequently exhibited in biological applications. They have been successfully applied to estimate metabolic fluxes,


Start
STEP 1 : Initialization
STEP 2 : PCA of "X"
STEP 3: PCA of "Y"
STEP 4: Obtain and Maximize the covariance between X & Y
Result
"Y"→ Input Dataset
"X"→ Output Dataset

(a)


X3
t1
Latent Variable
u1
Y3
X2
Y2
"PCA of X"
X1
PLS aim to maximize the covariance between X and Y
Y1
"PCA of Y"
u1
t1
"PLS Inner relation"

(b)


High Dimensional Dataset
Y1
Y2
Y3
Yj-2
Yj-1
Yj
Y4
PCA
Low Dimensional Dataset
X2
PC 2
PC 1
X1
PC → Principal Component

(c)

◂**Fig. 2** PLS algorithm **a** workflow, **b** purpose, and **c** working of PCA in association with the procedure of PLS algorithm

foresee membrane contamination in bioreactors, and find nucleotides for nanopore sensing. Bioprocess monitoring and control rely heavily on SVM estimates of important state variables, including product and biomass concentrations (Ge et al. 2023). Because of the time-varying nature of complex biochemical processes, researchers often use a learning-prediction approach for the construction of SVM training databases. SVM outperformed a few ANN algorithms (FBNN, RBFN, and RNN) in terms of efficiency and reliability in a study aimed at the estimation of product concentration in industrial penicillin manufacturing (Chen et al. 2015).

The application of SVM in predicting crucial state variables within bioprocesses, including product and biomass concentrations, has demonstrated notable success (Zhang et al. 2017). A rolling learning-prediction procedure designed to address the time-varying nature of the bioprocess has also been reported (Li and Yuan 2006). This method aids in the establishment of a comprehensive training database for the SVM predictor. SVM exhibited remarkable prediction accuracy, achieving an impressive range of 1–3% accuracy for predictions spanning 4–40 h into the future. Notably, this performance surpassed the best results obtained from three distinct ANNs examined.

Although SVM promises great results for the prediction of critical state variables in the realm of bioprocesses, many challenges remain to be resolved. Limitations associated with SVM algorithms are realized when this procedure is applied to large datasets. The reason could be the demand for large memory space by kernel functions. This is computationally expensive, i.e., "memory demands are $O(Y^2)$ and computation demands $O(Y^2 y_x)$ operations," where Y and y_x represent the number of data points and variables, respectively (Widodo and Yang 2007). In addition, SVMs do not always give a probabilistic output in the forecast they are making.

Artificial neural network (ANN)

ANNs are a class of ML models inspired by the structure and functioning of the human brain. ANN comprises interconnected nodes organized into layers: an input layer, hidden layers, and an output layer. Each connection is associated with a weight, and each node has an associated activation function (Lancashire et al. 2008). The mathematical representation involves a series of matrix multiplications and activation function applications during forward and backward propagation. The forward pass computes the predicted output, and the backward pass adjusts the weights using optimization algorithms like gradient descent and backpropagation to minimize the difference between predictions and actual outcomes (Phung 2019). The network learns by iteratively updating weights to enhance its ability to generalize from training data to unseen examples, making ANNs powerful tools for complex pattern recognition and nonlinear function approximation.

Predicting fermentation variables, characterizing amino acid sequences, simulating bioprocesses, biomass concentration estimation and modelling bioremediation of polluted groundwater are few applications of ANN in biochemical engineering (Natarajan et al. 2021). It takes into consideration many factors, like ionic strength and media composition. Thanks to a high number of adjustable parameters, neural networks are very general-purpose and may likely be trained to approximate any function. Parameters are often repeatedly optimized so that the neural network can become more efficient and produce accurate results.

There are different types of ANN, such as Deep Neural Network (DNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), etc. All these types of neural networks have found their application in the field of biochemical engineering (Krishna et al. 2021). CNN has been used for optimization of classification of enzyme tertiary structure in metabolic engineering, algal biofuel production, bioreactor optimization, and prediction of protein structure with incredible precision (Shin et al. 2016). RNNs find applications in genetic engineering for the recognition of promoter sequences, protein engineering for the statistical summarization of amino acid sequences, and biosensor calibration for soft sensors (Rivera and Karim 1992). The discovery that many hidden layers might ease the modelling of exceedingly complicated underlying functions due to the large number of parameters involved led to the creation of DNNs, which was possibly the most significant structural breakthrough. Figure 4 shows

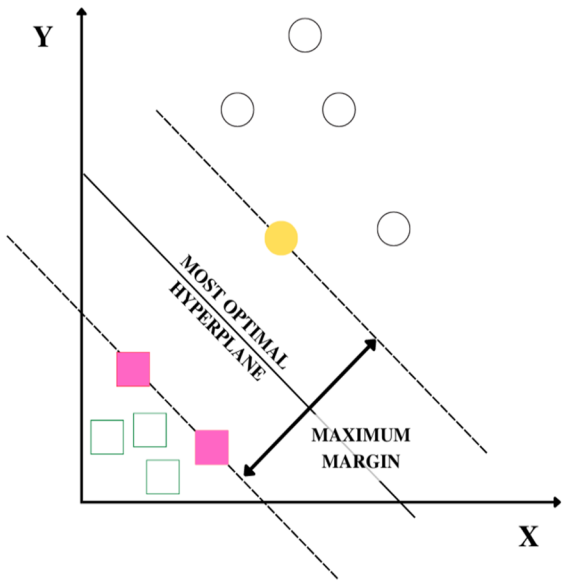


Fig. 3 Representation of support vector machines (SVM)

diagrammatic representations of workflows for ANN, CNN, RNN, and DNN.

Choosing the right number of hidden layers, neurons per hidden layer, and activation function is critical to the effectiveness of neural networks (Ren et al. 2017). Various methodologies have been used, ranging from the simple (such as a random search of hyper-parameters) to the complicated (such as Bayesian optimization) (Brochu et al. 2010). Unfortunately, the importance of hyper-parameter selection in neural network models is commonly underestimated. Choosing suitable hyperparameters is vital in neural network applications for feature selection and dimensionality reduction, like auto-encoders applied in cleaning RNA sequence databases and producing latent representations of genomic data.

Reinforcement learning (RL)

RL is a ML paradigm where an agent learns to make sequential decisions by interacting with an environment to maximize a cumulative reward signal. The learning technique uses a reward and penalty approach for reaching the best optimal result (Abbeel and Ng 2011; Popova et al. 2018). Unlike traditional approaches such as supervised and unsupervised learning, RL utilizes a trial-and-error methodology. The mathematical formulation of RL involves the Markov Decision Process (MDP), characterized by states, actions, transition probabilities, and rewards. The agent learns a policy, denoted as π, which maps states to actions, aiming to optimize the expected cumulative reward (Mahmud et al. 2018). The value function, represented by Q-values or state–action values, quantifies the long-term reward of taking a specific action in each state. The fundamental RL algorithms, such as Q-learning or policy gradient methods, employ mathematical mechanisms like Bellman equations, temporal difference updates, and stochastic gradient descent for learning optimal policies through trial and error. The convergence and effectiveness of RL models heavily rely on these mathematical foundations, enabling applications in the field of biochemical engineering (Petsagkourakis et al. 2020). Figure 5 gives a visual representation of the RL technique.

In the evolving landscape of bioprocess engineering, the integration of RL techniques, both model-based and model-free, represents a paradigm shift in how we approach process optimization and control. While model-based RL offers a structured and model-driven approach, model-free RL thrives in dynamic and uncertain environments (Pan et al. 2021). These methods provide the means to adapt and fine-tune bioprocesses with unprecedented precision and efficiency. Whether it is the production of life-saving pharmaceuticals, sustainable biofuels, or efficient wastewater treatment, the applications of RL are diverse and impactful.

A particular study presents a special RL algorithm called REINFORCE, for the optimization of "continuous control variables," "u" (Petsagkourakis et al. 2022). Often, Value-Based Algorithms are represented by Q-learning. They build the optimal policy by greedily picking the control with the greatest Q(x, u) value, which is learned and optimized directly from the state–action value function (state–action value). Actor–critic techniques are one kind of hybrid algorithm that combine policy optimization with value-based approaches. For instance, in the production of monoclonal antibodies, RL systems can dynamically adjust bioreactor conditions, resulting in a significant increase in yield while maintaining product quality. In bioethanol production, RL optimizes fermentation parameters, leading to higher yields and reduced energy consumption (Petsagkourakis

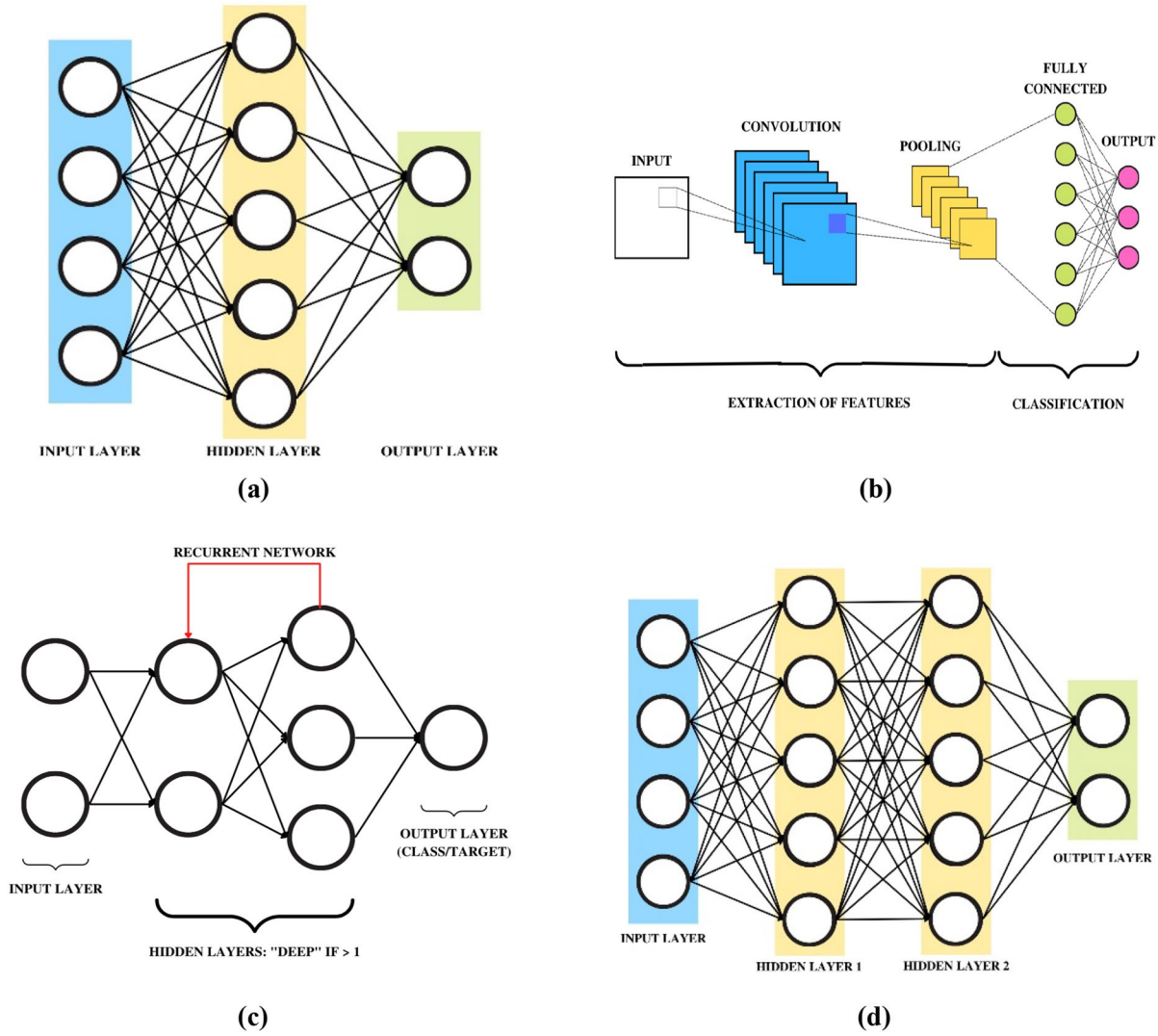


Fig. 4 Diagrammatic representations of workflows **a** Artificial Neural Network (ANN), **b** Convolutional Neural Network (CNN), **c** Recurrent Neural Network (RNN), and **d** Deep Neural Network (DNN)

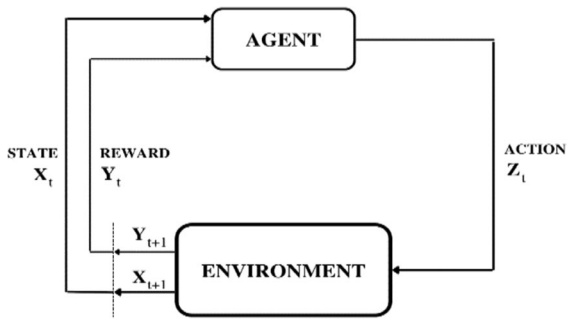


Fig. 5 Workflow of reinforcement learning algorithm

et al. 2022). In wastewater treatment, RL enhances the removal of contaminants and reduces chemical usage while meeting stringent effluent standards. RL is also involved in microbial biotechnology processes to yield more bio-based chemicals with cost-efficient production (Mowbray et al. 2022). Furthermore, for bioreactor control, RL outperforms traditional controllers, achieving precise regulation and minimizing batch-to-batch variability (Duong-Trung et al. 2023). These case studies underscore the transformative impact of RL in the field of biochemical and bioprocess engineering, offering not only operational

efficiency but also sustainability, cost-effectiveness, and the promise of further innovations in the industry (Nian et al. 2020).

Applications of ML algorithms in bioprocess engineering

ML algorithms find application in diverse sub-fields of bioprocess engineering. These include bioreactor optimization, bioreactor control, microbial co-culture optimization, fermentation process optimization, batch bioprocess optimization, metabolic engineering, bioprocess modelling, and bioresources and bioenergy processes, etc.

Bioreactor optimization

Bioprocess development, an integral part of the biopharmaceutical industry, remains a challenging and resource-intensive endeavor (Smiatek et al. 2020). Among its intricacies, optimizing the culture media used for bioprocessing can be particularly time-consuming and labour-intensive. In a study on the generation of therapeutic proteins, researchers explored the application of Bayesian optimization (BO) models (Alcântara 2018). The primary focus of this investigation was the enhancement of the feeding strategy within a pre-existing fed-batch protocol for Chinese Hamster Ovary (CHO) cell lines, specifically those involved in the production of mAb derived proteins. To carry out this ambitious project, BO and its various iterations were put to the test using both *in silico* and experimental models. The researchers initiated the optimization process by employing computer simulations, leveraging an *in silico* model that mirrored a fed-batch process (Cougnon et al. 2011). Notably, one study within this research adopted an approach involving Latin hypercube sampling for initialization and employed expected improvement as an acquisition function. Additionally, local penalization was integrated into the BO methodology (Sampaio and Calado 2019; Viana 2015). This was achieved through a series of systematic experiments to determine the most effective feeding strategy among four different options, ultimately yielding the highest protein titres. The optimization process began with an initialization phase comprising 16 tests, followed by four batches of experiments, each conducted in 50 mL spin tubes. Impressively, this comprehensive study resulted in a remarkable 35% increase in protein titres compared to the baseline method. To further validate the applicability of the scaled-down models, the research extended into 3 L benchtop bioreactors, reaffirming the promising results. Moreover, this study delved into the development and deployment of several ML algorithms, all meticulously implemented using MATLAB functions. Key algorithms such as initialization methods, surrogate models, and acquisition functions were precisely employed, further enhancing the efficiency and accuracy of the optimization process.

The insights garnered from this study open exciting prospects for the optimization of feeding strategies, not only for bioprocesses but also for the formulation of ideal culture media in perfusion cultures. It also underscores the importance of addressing concerns related to medium osmolarity. In a larger context, these findings pave the way for the discovery of optimal operating parameters through the utilization of perfusion scale-down models, promising advancements in bioprocessing efficiency, and the production of therapeutic proteins (Pereira et al. 2018; Zheng et al. 2017).

Bioreactor control

In recent scientific investigations, there has been a growing interest in the application of RL and ANN for the control of complex industrial processes. Traditional inverse model ANN-based control methods exhibit remarkable adaptability to intricate nonlinear systems, even with limited training data, while RL algorithms introduce a distinct approach. They establish an optimal control policy through iterative exploratory operations, albeit at a somewhat slower convergence rate.

Notably, a groundbreaking strategy has emerged to circumvent the need for extensive training data in the context of inverse ANN-based control. This innovative approach, known as Partially Supervised Reinforcement Learning (PSRL), leverages ANNs to acquire an initial control policy. PSRL represents a compelling solution for addressing the significant economic challenges associated with regulating yeast fermentation within a semi-continuous fed-batch

bioreactor (Xu 2019). This control problem is formally framed as a 'Markov Decision Process (MDP)', utilizing a synthesis of RL and PSRL techniques.

The bioreactor control problem can be formulated as MDP with a state space S, an action space A, a transition function T(s, a, s') for all $s \in S$ and $a \in A$, a reward function R(s, a), and a discount factor γ. The goal is to learn a policy $\pi(a|s)$ that maximizes the expected cumulative reward over time. Through comprehensive comparative assessments involving model-based and model-free RL control in practical tests and simulations, PSRL consistently emerges as the superior performer, excelling in a diverse array of performance criteria when juxtaposed with RL and inverse model ANN-based control techniques.

Microbial co-culture optimization

Deep RL techniques have found applications in the control of microbial co-cultures as well, where they are employed to optimize growth conditions, regulate metabolic pathways, and manage community dynamics for various biotechnological purposes, such as biofuel production, bioremediation, and pharmaceutical development. Treloar et al. (2020) harnessed the power of Neural Fitted Q-Learning to control a model representing microbial co-culture. This study demonstrated that RL surpasses traditional industrial control systems, particularly when dealing with extended sample-and-hold intervals. RL is especially advantageous when dealing with situations where data limitations are not a constraint. Furthermore, investigations have shown that RL can efficiently optimize microbial co-cultures in cost-effective bioreactors. This implies that it is possible to optimize the output of a microbial population without the need for a specialized model, relying instead on information about microbial population levels and the pace at which the desired product is generated. The approach introduces a specific optimization methodology employing a chemostat model for a bioreactor. "Auxotrophy" was utilized as a control mechanism, although other control measures such as "independent carbon sources," "induced lysis," or "growth arrest" could also be implemented. Notably, controlling microbial populations can potentially lead to mutations. Therefore, the RL technique utilized in this study involves the continuous updating of the agent's policy, ensuring flexibility to adapt to evolving growth dynamics.

Given the iterative nature of RL, only a few studies have explored techniques that may influence evolutionary changes in growth dynamics. While RL offers a promising avenue, model-based alternatives such as dynamic programming might also contribute to optimal control challenges.

Researchers have underscored the efficacy of deep RL in managing diverse ecosystems. The significance of community engineering is expected to grow in the context of synthetic biology and industrial biotechnology as production processes become increasingly complex (Fisher et al. 2022). As a result, the application of advanced AI techniques to steer intricate biological systems is becoming increasingly valuable (Bain and Sammut 2000).

Fermentation process optimization

Fermentation technology plays a crucial role in various applications, and the ability to optimize fermentation conditions is of paramount importance. A study specifically focused on the enhancement of Iturin A production, employed various approaches, including ANNs and SVMs (Chen et al. 2015). Independent variables such as the concentrations of asparagine, glutamic acid, and proline (mg/L) were meticulously considered, while the dependent variable was the "Iturin A titer" (Tang et al. 2021). Multiple ML models, including SVM, multilayer feed-forward neural networks (MLFNs), and general regression neural networks (GRNNs), were developed and meticulously compared to determine their performance in this context. Notably, MLFNs exhibited higher root-mean-square error, longer training periods, and significant variability with shifts in the number of nodes. In contrast, GRNNs showcased the lowest root-mean-square error and the quickest training period, maintaining consistent performance across various experiments. The SVM model also demonstrated a brief training period and a minimal root-mean-square error. The model findings highlight that the GRNN excels as the preferred choice for establishing fed-batch fermentation conditions aimed at enhancing Iturin A production. Renowned for its robustness and high precision, the GRNN emerges as the top performer. The SVM stands as a viable alternative model with comparable

performance. Both the GRNN and SVM consistently achieve prediction accuracies exceeding 100% within a 30% tolerance margin across independent experiments.

Thus, it can be concluded that the application of ML, particularly through the GRNN and SVM models, provides a potent means to optimize fermentation processes, delivering accuracy and robustness essential to produce compounds like Iturin A.

Batch bioprocess optimization

Many scientists are looking for bioprocesses to replace their use of fossil fuels with more environmentally friendly options. However, they might be difficult to optimize because of their non-stationary operation modes and stochastic behaviour (Duong-Trung et al. 2023). Plant model mismatch is very widespread since biological systems are so complicated. An optimization strategy for batch processes, based on RL, is one option for dealing with the aforementioned problems.

The research work carried out by Petsagkourakis et al. (2019), presents a novel approach for optimizing bioprocesses utilizing RL techniques. By leveraging RL algorithms, particularly the "REINFORCE" algorithm, the study addresses the inherent challenges posed by the stochastic nature of bioprocess dynamics and the discrepancies between plant models and real-world systems. In the first stage of the methodology, an initial control policy is developed offline using an approximate model of the bioprocess system. Through extensive training epochs, the policy network, represented by a RNN, learns to optimize control actions to maximize a predefined reward function, such as the concentration of target products. Subsequently, in the second stage, the control policy undergoes refinement using real plant data in a batch-to-batch framework. This real-world data enables online updates to the policy network, enhancing its accuracy and adaptability to actual bioprocess conditions. The stochastic control policy, implemented through the RNN, dynamically predicts optimal control actions based on current states, past actions, and remaining batch time. By iteratively refining the control policy through a combination of offline training with approximate models and online updates with real plant data, the proposed methodology demonstrates significant advancements in bioprocess optimization under uncertainty. This research offers valuable insights and practical implications for improving the efficiency and reliability of bioproduction systems in various industrial applications.

Consequently, the utilization of RL, specifically the batch-to-batch policy gradient approach, exhibits great potential in the optimization of complex bioprocesses. This approach offers a proactive strategy for addressing the challenges posed by non-stationary modes and system complexities, thereby contributing to the development of sustainable energy solutions.

Metabolic engineering

ML also finds applications in the field of metabolic engineering. Here, two different case studies in the field are discussed.

Predicting metabolic circuit kinetics

Anticipating the kinetics of a metabolic circuit is a common challenge in bioengineering that can be effectively addressed through ML. Pathway dynamics prediction substantially enhances the efficiency of the pathway design process by allowing us to foresee which pathway designs will meet our requirements in advance (Narayanan et al. 2021). Classical kinetic models, with an explicit functional link between substrate/product concentrations and enzymes, have traditionally been employed to estimate the rate of change of specific metabolites. The Michaelis–Menten kinetic model has been the gold standard for investigating enzyme metabolism over time. However, for many processes, the precise functional link between metabolites and enzymes remains elusive, leading to suboptimal predictive ability due to knowledge gaps regarding the underlying mechanisms.

Costello and Martin (2018) introduced an innovative approach to predict metabolite dynamics by utilizing supervised ML, focusing on the limonene/isopentenol ratio. The core methodology involved developing a function based on protein and metabolite concentration data, prioritizing predictive accuracy over an intricate understanding of underlying biochemical processes. To enhance the model's performance, the researchers conscientiously grappled with the challenge of dimensionality, recognizing the

potential benefits of additional data while considering the exponential growth in data requirements, known as the 'curse of dimensionality.' A pivotal aspect of their strategy was the employment of feature selection algorithms, systematically identifying and retaining the most influential input features. This process not only streamlined the model architecture but also addressed issues of redundancy and overfitting. Synthetic features, derived from the original dataset, were introduced intelligently to strike a balance between model simplicity and accuracy. Additionally, dimensionality reduction techniques were judiciously applied to manage the complexity of the dataset while preserving predictive precision. The integration of a tree-based pipeline from scikit-learn ML models brought automation to the feature preprocessing approach, further enhancing efficiency and objectivity. This automated strategy not only expedited the feature selection process but also ensured a systematic and data-driven decision-making process. Hence, their study exemplifies a nuanced integration of ML and feature engineering strategies to optimize predictive accuracy while navigating the challenges posed by high-dimensional datasets.

Enhancing *n*-butanol production pathways

Metabolic pathways serve as intricate networks guiding biochemical transformations, crucial for synthesizing essential molecules and driving cellular functions (Oberhardt et al. 2009). The complexity of metabolic pathways, such as those involved in *n*-butanol production, underscores the need for innovative approaches like ML to unravel intricate relationships and optimize pathway design efficiently. In recent years, statistical techniques like PCA have emerged as invaluable tools for analyzing data derived from engineered organisms, offering insights crucial for guiding subsequent rounds of design. Historically, the complexity of data obtained from proteomics analysis posed challenges in discerning meaningful trends for system understanding and reengineering. However, advancements in techniques such as PCA have revolutionized this landscape by enabling the analysis of small datasets to unveil patterns or trends essential for informing system redesign. As datasets become increasingly abundant and diverse, the necessity for more sophisticated data-analysis tools, including ML algorithms, becomes apparent (Nielsen and Keasling 2016; van Lent et al. 2023).

The study conducted by Karim et al. (2020) provides a compelling demonstration of how ML techniques revolutionized the optimization of *n*-butanol production pathways. Central to the exploration is the elucidation of *in vitro* Prototyping and Rapid Optimization of Biosynthetic Enzymes" (iPROBE) methodology, which involves utilizing cell lysates enriched with biosynthetic enzymes through cell-free protein synthesis. This process allows for the rapid assembly and testing of metabolic pathways, facilitating the assessment of pathway performance *in vitro* (Batista et al. 2021).

Through a series of experiments, the researchers showcased the platform's capability to drastically reduce the time required for pathway construction while enabling the exploration of a vast array of enzyme combinations. By employing cell-free prototyping within the iPROBE framework, they were able to accelerate the design-build-test-learn (DBTL) cycles in metabolic engineering from months to mere days. The results were impressive, with ML-guided pathway designs significantly outperforming traditional base-case pathways, achieving over fourfold improvements in butanol production scores. However, the study also shed light on the challenge posed by the immense number of potential pathway combinations, emphasizing the need for a data-driven approach to pathway design. Through the utilization of DNNs, the researchers effectively navigated this complexity, generating optimized pathway designs that surpassed expert-determined alternatives. The neural network crafted in this case featured "5–15 fully connected hidden layers" and "5–15 nodes per layer." Training the neural network relied on a dataset comprising 120 pathway designs, with genetic algorithms assisting in determining the optimal neural network architecture. Looking ahead, the scalability of ML models in metabolic engineering hinges on the availability of abundant, high-quality, and well-organized data. While challenges such as data scarcity and dimensionality remain, continued advancements in ML techniques offer promising solutions to overcome these obstacles.

In conclusion, this study underscores the transformative potential of ML-guided pathway optimization in bioprocess engineering. By leveraging the power of computational methods within innovative

platforms like iPROBE, researchers can accelerate the development of sustainable bio-based production processes, paving the way for a more efficient and environmentally friendly future in industrial biotechnology. These studies have ushered in a new era of pathway optimization characterized by expeditiousness and have harnessed the cutting-edge capabilities of neural network predictions.

Bioprocess modelling

ML technology has also been put to good use in modelling and optimizing bioprocesses (Fakharudin et al. 2007). Among the various ML techniques, gradient boosting stands out as a unique ensemble learning strategy that combines weak learners to yield enhanced predictions. Despite its significant promise in improving the dynamic modelling and prediction of bioprocesses, gradient boosting remains a relatively understudied avenue.

Some researchers have initiated preliminary investigations into gradient boosting, particularly in its capacity to outperform commonly used ANNs. This exploration arises from the inherent challenge of generating a substantial volume of experimental data for model training in many bioprocesses. Researchers have demonstrated through diverse experimental verifications that gradient boosting surpasses ANNs in predictive modelling for bioprocesses, positioning it as a compelling choice for the future of bioprocess digitalization and optimization.

Moreover, ML models, acting as surrogate models, provide a practical alternative to intricate scale-up models, such as substituting costly simulations in computational fluid dynamics, thereby offering a cost-effective and efficient solution for optimizing bioprocesses. Surrogate models can be built using ML algorithms like ANNs, SVMs, or decision trees. These models can be trained on data from laboratory-scale experiments and utilized to forecast bioprocess performance at larger scales.

Another application of ML in bioprocess modeling involves integrating mechanistic models with data-driven models. Mechanistic models delineate the fundamental physical and chemical processes occurring in a bioprocess, whereas data-driven models utilize statistical techniques to discern patterns in experimental data. By amalgamating these two model types, it becomes feasible to uncover non-traditional scale-up criteria by correlating validated mechanistic models that describe bioprocess performance across laboratory and production scales.

The equations utilized in ML models for bioprocess modeling may vary depending on the specific algorithm employed. For instance, ANNs leverage interconnected nodes to process input data and generate output predictions, with equations based on the activation functions used in the nodes. SVM employs a kernel function to transform input data into a higher-dimensional space, facilitating easier separation into different classes, with equations grounded in the specific kernel function used.

Therefore, ML holds considerable promise in enhancing bioprocess modeling by enabling efficient exploration of extensive design spaces and uncovering non-traditional scale-up criteria. Nonetheless, the existing literature on ML for bioprocess modeling remains limited, necessitating further development in this domain.

Bioresources and bioenergy

The negative environmental impact of fossil fuels has driven research into renewable alternatives, such as biofuels and bioenergy, to ensure a sustainable bioeconomy. Biofuel and bioenergy processes are vital components of this transition, necessitating modelling and optimization (Duong-Trung 2023). ML and deep learning techniques hold significant promise for creating reliable models in biofuel production (Bennamoun 2022). However, modelling operational parameters, such as biofuel extraction rate and operating temperature, poses a challenging task.

As biofuels offer potential benefits, including reduced emissions, decreased fuel consumption, and expanded engine operating ranges, they have become a focus of research in combustion engines. However, assessing engine performance with various biofuel blends is time-consuming and requires engine adjustments. ANNs are emerging as a solution to these challenges, enabling the modelling of biodiesel fuel blends in engines and predictions of performance indicators based on engine speed and biofuel percentage. Extending beyond traditional ANNs, Extreme Learning Machines (ELMs) have been explored to address this issue. ELMs offer faster learning and

improved generalization compared to standard ANNs (Wang et al. 2022). A variant called "Kernel-based Extreme Learning Machine (K-ELM)" has been introduced to predict gasoline engine performance. K-ELM demonstrates comparable or superior accuracy to the least-squares support vector machine (LS-SVM), while significantly reducing training and testing times.

Other applications in the field include predicting biogas and methane production during the anaerobic treatment of wastewater molasses using ANN models. ANN-based models are also employed to forecast the impact of operational factors (temperature, moisture content, and ash content) on hydrogen output in biomass gasification. Decision tree-based classifiers aid in determining optimal operational factors for algal biomass growth and lipid synthesis, with a focus on dividing training data into high and low classes based on specific criteria.

Other recent advancements in bioprocess engineering

In recent years, the field of bioprocess engineering has witnessed significant advancements fuelled by the integration of ML, deep learning, and AI techniques. These innovative technologies, exemplified by tools like chatbot, GPT-4, natural language processing (NLP) have revolutionized various aspects of biochemical and bioprocess engineering, offering novel solutions to complex challenges, and unlocking new opportunities for sustainable development (Heras et al. 2020; Xiao et al. 2023; Yang et al. 2023). As ML continues to evolve, its applications in bioprocess engineering have expanded beyond traditional boundaries, encompassing diverse areas such as metabolic pathway design, renewable energy production, waste management, and pharmaceutical research and development. The integration of generative AI, through the utilization of natural language processing tools like GPT-4, has revolutionized knowledge extraction from synthetic biology journal articles, streamlining information retrieval and empowering advancements in bioprocess engineering. Table 1 presents recent studies on the applications of ML, deep learning, and AI in the different sub-domains of biochemical and bioprocess engineering.

Challenges for the applications of ML in bioprocess engineering

Despite the significant advancements that ML has brought to bioprocess engineering, certain challenges persist, hindering its widespread adoption. Bioprocess engineers carefully select the most suitable ML models by comparing their performance on similar datasets and research problems. Benchmarking has been extensively explored as a method to assess and apply ML techniques to bioprocess datasets (Dhar and Shamir 2021; Takei and Ishida 2022). The utilization of ML techniques in bioprocess engineering holds immense promise for optimizing processes, enhancing efficiency, and driving innovation.

In Table 2, we outline the challenges faced in integrating ML into bioprocess engineering and propose practical solutions to address them. By addressing issues related to reproducibility, data quality, interpretability, and model robustness, the objective is to facilitate the seamless integration of ML into bioprocess engineering workflows. This approach aims to foster advancements in this interdisciplinary field.

Conclusion

In the domain of bioprocess engineering, the utilization of ML techniques is driven by three primary objectives: facilitating information retrieval, predicting dynamic responses, and reducing overall effort. It is imperative to emphasize that, in the context of biochemical engineering studies, the knowledge underpinning the information gleaned from data-driven ML methods holds more significance than the data itself. The effectiveness of an ML model ultimately hinges on the tangible impacts of its predictions when applied to bioprocess data. Our exploration into previously undiscovered molecular pathways has revealed that ML models, in isolation, may not always suffice. The central aim is to provide reliable models for the biochemical engineering community, making comprehensive documentation a critical requirement.

While numerous data-driven modelling methods are employed for bioprocess control, optimization, and monitoring, caution must be exercised in the wholesale adoption of deep learning methodologies without adequate explanation (Fisher et al. 2020;

Table 1 Recent advancements of ML in the field of bioprocess engineering

S. No	Sub-field of the ML-based study	Outcome of the study	References
1	Design of metabolic pathways, accelerating the development of microbial cell factories	• The study aims at evaluating the use of digital strategies, including computer-supported tools and deep learning techniques for pathway prediction and enzyme discovery, which streamline the design process of microbial cell factories, and bolsters the prospects of a bio-based circular economy	Ryu et al. (2023)
2	Optimization of gasification for efficient hydrogen production	• The study fosters green economy and circular energy practices, outlining a path towards a sustainable, pollution-free future	Bin Abu Sofian et al. (2024)
3	Optimization of anaerobic co-digestion (AcoD)	• The study provides a comprehensive understanding of operational parameters • Understanding the challenges of manual monitoring and introduction of AI as an innovative solution for AcoD process optimization, control, prediction, and real-time monitoring	Khan et al. (2023)
4	Natural language processing to review the literature on AI in bioprocessing	• By analysing papers from 2013 to 2022, focusing on keywords related to bioprocessing and AI, the study identifies a significant concentration (50%) on hybrid models, ANN, biopharmaceutical manufacturing, and biorefinery in the last 5 years • The findings highlight the potential of AI, particularly in enhancing design and process engineering strategies within the bioprocessing domain	Yang et al. (2023)
5	Application of generative models in protein design, emphasizing their potential for creating tailored sequences for bioenergy applications and the importance of quality datasets	• The review highlights generative ML models' effectiveness in rapid protein sequence generation for bioenergy, addressing energy production hurdles • Moreover, the study elucidates how these ML-based techniques aid in optimizing strains to maximize bioenergy yields, thus contributing significantly to sustainable energy solutions	Barghout et al. (2023)
6	Proposes integration of a chatbot into domain of bioprocess engineering	• This study proposes a chatbot, designed as a "friend" or colleague, which enhances collaborative learning on digital platforms • The chatbot is embedded in an open-source Educational Virtual Bioprocess Plant, FermProc, providing a supportive and interactive learning environment for researchers in bioprocess technology	Heras et al. (2020)
7	Integration of AI based tools within pharmaceutical R&D	• This study aims at introducing a comprehensive platform for analyzing histopathology data in the context of pharmaceutical R&D • It utilizes cutting-edge AI technology integrated into an open-source image viewing platform, enabling clinicians, to effortlessly utilize pre-trained algorithms	Lutnick et al. (2023)

Table 1 (continued)

S. No	Sub-field of the ML-based study	Outcome of the study	References
		• The system facilitates collaboration across disciplines and supports a range of preclinical research tasks, including tissue segmentation and patient-level prediction • Furthermore, it provides flexibility for deploying and evaluating new models, thus enhancing its applicability within pharmaceutical workflows	
8	Development of an advanced combinatorial ANN design-of-experiment (ANN-DoE) model tailored for optimizing parameters of mRNA-LNP (lipid nanoparticles) vaccines	• This study utilized a definitive screening design and ML algorithms to systematically evaluate key factors like ionizable lipid, N/P ratio, and flow rate ratios • Notably, the ANN-DoE model exhibits superior predictive accuracy compared to independent ML models, showcasing its effectiveness in refining critical vaccine parameters • This comprehensive approach implies a significant stride in mRNA-LNP vaccine development, holding promise for enhanced vaccine efficacy and accelerated optimization processes	Maharjan et al. (2023)

Emmert-Streib et al. 2020). Although ML and deep learning models have demonstrated superior accuracy and performance compared to physical models, these metrics should not be the sole criteria for evaluating their effectiveness, creativity, and relevance in the context of bioprocess engineering (Gupta et al. 2021).

It is crucial to remember two fundamental principles when delving into the regulation and optimization of bioprocesses using ML models. First, it is essential to acknowledge that "all models are flawed, but some are valuable." Second, simplicity often suffices when constructing models, as complexity or scale does not inherently imply strength or value. While these methods can prove highly effective when appropriately applied, we must recognize that they are not infallible solutions. Attempts to enhance the fit of a complex ML model by adjusting or increasing the number of parameters may seem appealing, but they can jeopardize the quality of research and diminish its potential impact. Contrary to the notion of a single optimal approach to ML, it is crucial to recognize that each approach has its own merits and limitations. Therefore, when implementing ML methodologies, a deep understanding of the underlying biochemical processes must be maintained throughout the study.

ML techniques have the potential to address the challenges encountered in mechanistic modelling. The data-driven nature of ML, coupled with robust performance metrics, allows for the extraction of highly accurate results, particularly in the realm of biochemical process engineering. This may manifest as the optimization of pathway designs, the enhancement of scale-up processes, or the improvement of bioreactor control. However, it is imperative to acknowledge the limitations of ML, including the demand for substantial amounts of data and challenges related to the interpretation of complex neural networks like DNNs.

Moreover, ML also finds application in the field of metabolic engineering, although it has received relatively less attention compared to the optimization of bioengineering processes. A few papers have delved into this area, with some studies breaking down the application of ML in metabolic engineering into stages such as "gene annotation and pathway design, pathway optimization, pathway construction, performance testing, and production scale-up." Furthermore, new ML techniques, like transfer learning, are continually emerging, but their specific applications within this field remain relatively unexplored.

Table 2 Major challenges for the applications of ML in various fields of biochemical engineering and their potential solutions

Challenges	Problems	Potential solutions
Reproducibility in ML	• Lack of comprehensive documentation, including code, data, and parameter settings, hampers reproducibility • Difficulty in replicating experiments due to variations in implementation and settings	• Encourage authors to provide detailed documentation, including code, datasets, and parameter settings • Establish repositories for open-access code and data to facilitate reproducibility
Complexity in Biological Systems	• Challenges in reproducing experiments due to variability in living organisms, chemicals, and experimental conditions • Lack of "FAIR (Findable, Accessible, Interoperable, Reusable)" data and difficulties in replicating model fitting strategies	• Develop and utilize robust statistical methods capable of handling variability and uncertainty inherent in biological systems • Techniques such as Bayesian inference and hierarchical modeling can account for data heterogeneity and provide more reliable estimates
Limitations of ML Platforms	• Inability of ML platforms to ensure computational reproducibility • Test results generated from these platforms may not always be trustworthy	• Foster open-source collaboration and peer review within the ML community to address platform limitations and improve computational reproducibility • Shared code repositories, community forums, and collaborative projects can facilitate knowledge sharing and continuous improvement of ML platforms
Data Quality and Quantity	• Bioprocess engineering often deals with heterogeneous, noisy, and sparse data, which can affect the performance of ML models • Insufficient data quantity or poor data quality may lead to biased or unreliable predictions	• Employ data augmentation techniques such as varying fermentation conditions, altering nutrient composition, adjusting pH levels, and changing agitation rates to increase the quantity and diversity of training data for bioprocess models, enhancing their robustness and predictive capabilities • Additionally, invest in robust data preprocessing pipelines to handle noise, missing values, and outliers, ensuring high-quality input data for ML models
Interpretability and Explainability	• The interpretability of ML models, particularly complex ones like deep learning, remains a challenge • Understanding the reasoning behind model predictions is crucial for trust and adoption, especially in critical applications	• Tree-based ML algorithms, such as random forest and gradient boosting, offer better interpretability compared to deep learning models • Emphasizing the use of interpretable algorithms, such as decision trees, logistic regression, and linear regression, can enhance the trustworthiness of ML-based predictions by providing clear insights into how inputs are being transformed into outputs
Incorporating Biological Data	• Conventional models often overlook the rich biological data regarding microbial communities, leading to incomplete representations of bioprocesses • Integrating biological information into ML models poses a challenge due to the complexity and high dimensionality of biological data	• Develop methodologies to incorporate microbial community data into ML models to provide a more comprehensive understanding of bioprocesses • Techniques like feature engineering and transfer learning can help extract meaningful patterns from complex biological data

Table 2 (continued)

Challenges	Problems	Potential solutions
Model Robustness and Generalization	• Ensuring the robustness and generalization capability of ML models across different experimental setups, environmental conditions, and datasets is challenging • Models trained on specific datasets may struggle to perform well on unseen data or under varying conditions	• Implement regularization techniques such as dropout and L1/L2 regularization to prevent overfitting and improve model generalization • Ensemble learning methods, such as bagging and boosting, can also enhance model robustness by combining multiple models' predictions

By understanding the advantages and limitations of each approach, bioprocess engineers can make informed decisions, harnessing the power of ML to drive innovation and excellence in bioprocess engineering. The future holds exciting possibilities for the synergy of ML and bioprocess engineering, promising greater sustainability, productivity, and reliability in the bioprocessing industry.

Acknowledgements The authors are grateful to the Delhi Technological University, New Delhi, India, for providing an opportunity to conduct this study.

Author contributions All authors contributed to the study conception and design. The first draft of the manuscript was written by Ritika Saha and Ashutosh Chauhan and all authors commented on previous versions of the manuscript. The manuscript was critically reviewed by Smita Rastogi Verma. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability No additional data available.

Consent to publish No human or animal subjects were involved.

Declarations

Conflict of interest There is no conflict of interest. The authors have no relevant financial or non-financial interests to disclose.

Ethical approval No human or animal subjects were involved.

Informed consent No human or animal subjects were involved.

References

Ab Azar N, Shahmansoorian A, Davoudi M (2020) From inverse optimal control to inverse reinforcement learning: a historical review. Annu Rev Control 50:119–138. https://doi.org/10.1016/j.arcontrol.2020.06.001

Abbeel P, Ng AY (2011) Inverse reinforcement learning. In: Encyclopedia of machine learning. Springer, Boston, pp 554–558. https://doi.org/10.1007/978-0-387-30164-8_417

Abdi H, Williams LJ (2010) Principal component analysis. WIREs. Comput Stat 2(4):433–459. https://doi.org/10.1002/wics.101

Alcântara JMG (2018) Machine learning based optimization of bioprocesses for the production of therapeutic proteins. Thesis, Universidade de Lisboa

Bain M, Sammut C (2000) A framework for behavioural cloning. Mach Intell 15:103–129. https://doi.org/10.1093/oso/9780198538677.003.0006

Barghout RA, Xu Z, Betala S, Mahadevan R (2023) Advances in generative modeling methods and datasets to design novel enzymes for renewable chemicals and fuels. Curr Opin Biotechnol 84:103007. https://doi.org/10.1016/j.copbio.2023.103007

Batista AC, Soudier P, Kushwaha M, Faulon J (2021) Optimizing protein synthesis in cell-free systems, a review. Eng Biol 5(1):10–19. https://doi.org/10.1049/enb2.12004

Bennamoun L (2022) Bioresource technology for bioenergy: development and trends. Energies 15(5):1717. https://doi.org/10.3390/en15051717

Bhange K, Nath A, Singh N, Chaturvedi V, Bhatt R (2023) Statistical optimization and prediction of significant nutritional factors for keratinase production by *Stenotrophomonas maltophilia* Kb2 and its application as dehairing agent. Bioresour Technol Rep 23:101541. https://doi.org/10.1016/j.biteb.2023.101541

Bin Abu Sofian ADA, Lim HR, Chew KW, Khoo KS, Tan IS, Ma Z, Show PL (2024) Hydrogen production and pollution mitigation: enhanced gasification of plastic waste and biomass with machine learning and storage for a sustainable future. Environ Pollut 342:123024. https://doi.org/10.1016/j.envpol.2023.123024

Boulesteix A-L, Strimmer K (2006) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform 8(1):32–44. https://doi.org/10.1093/bib/bbl016

Brochu E, Cora VM, De Freitas N (2010) A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint. https://doi.org/10.48550/arXiv.1012.2599

Cantarero-Rivera FJ, Yang R, Li H, Qi H, Chen J (2024) An artificial neural network-based machine learning approach to correct coarse-mesh-induced error in computational fluid dynamics modeling of cell culture bioreactor. Food Bioprod Process 143:128–142. https://doi.org/10.1016/j.fbp.2023.11.004

Chen F, Li H, Xu Z, Hou S, Yang D (2015) User-friendly optimization approach of fed-batch fermentation conditions for the production of iturin A using artificial neural networks and support vector machine. Electron J Biotechnol 18(4):273–280. https://doi.org/10.1016/j.ejbt.2015.05.001

Costello Z, Martin HG (2018) A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. NPJ Syst Biol Appl. https://doi.org/10.1038/s41540-018-0054-3

Cougnon P, Dochain D, Guay M, Perrier M (2011) Online optimization of fedbatch bioreactors by adaptive extremum seeking control. J Process Control 21(10):1526–1532. https://doi.org/10.1016/j.jprocont.2011.05.004

Cutajar K, Pullin M, Damianou A, Lawrence N, González J (2019) Deep Gaussian processes for multi-fidelity modeling. arXiv preprint. https://doi.org/10.48550/arXiv.1903.07320

David AN, Sewsynker-Sukai Y, Meyer EL, Kana EBG (2023) Harnessing artificial neural networks and large language models for bioprocess optimization: predicting sugar output from kraft waste-based lignocellulosic pretreatments. Ind Crops Prod 206:117686. https://doi.org/10.1016/j.indcrop.2023.117686

Dhar S, Shamir L (2021) Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks. Vis Inform 5(3):92–101. https://doi.org/10.1016/j.visinf.2021.10.001

Dürrenmatt DJ, Gujer W (2011) Data-driven modeling approaches to support wastewater treatment plant operation. Environ Model Softw. https://doi.org/10.1016/j.envsoft.2011.11.007

Duong-Trung N, Born S, Kim JW, Schermeyer MT, Paulick K, Borisyak M, Cruz-Bournazou MN, Werner T, Scholz R, Schmidt-Thieme L, Neubauer P, Martinez E (2023) When bioprocess engineering meets machine learning: A survey from the perspective of automated bioprocess development. Biochem Eng J. 190:108764. https://doi.org/10.1016/j.bej.2022.108764

Emmert-Streib F, Yang Z, Feng H, Tripathi S, Dehmer M (2020) An introductory review of deep learning for prediction models with big data. Front Artif Intell. https://doi.org/10.3389/frai.2020.00004

Esfandyari M, Fanaei MA, Gheshlaghi R, Akhavan Mahdavi M (2017) Mathematical modeling of two-chamber batch microbial fuel cell with pure culture of *Shewanella*. Chem Eng Res Des 117:34–42. https://doi.org/10.1016/j.cherd.2016.09.016

Fakharudin AS, Embong A, Hamid RA, Mohamad Hamza MA, Ajid KA, Ali NM, Satari SZ, Sulaiman J, Wan Md Zain WS (2007) Optimization of fermentation process using data mining techniques for small-medium industry. In: 2007 International conference on intelligent and advanced systems, 2007. https://doi.org/10.1109/icias.2007.4658389

Fisher OJ, Watson NJ, Escrig JE, Witt R, Porcu L, Bacon D, Rigley M, Gomes RL (2020) Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems. Comput Chem Eng 140:106881. https://doi.org/10.1016/j.compchemeng.2020.106881

Fisher OJ, Watson NJ, Porcu L, Bacon D, Rigley M, Gomes RL (2022) Data-driven modelling for resource recovery: data volume, variability, and visualization for an industrial bioprocess. Biochem Eng J 185:108499. https://doi.org/10.1016/j.bej.2022.108499

Ge Z, Song Z, Ding SX, Huang B (2017) Data mining and analytics in the process industry: the role of machine learning. IEEE Access 5:20590–20616. https://doi.org/10.1109/access.2017.2756872

Ge H, Zheng J, Xu H (2023) Advances in machine learning for high value-added applications of lignocellulosic biomass. Bioresour Technol 369:128481. https://doi.org/10.1016/j.biortech.2022.128481

Goodacre R (2003) Explanatory analysis of spectroscopic data using machine learning of simple, interpretable rules. Vib Spectrosc 32(1):33–45. https://doi.org/10.1016/s0924-2031(03)00045-6

Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P (2021) Artificial intelligence to deep learning: machine intelligence approach for drug discovery.

Mol Divers 25(3):1315–1360. https://doi.org/10.1007/s11030-021-10217-3

Helmy M, Elhalis H, Liu Y, Chow Y, Selvarajoo K (2023) Perspective: multiomics and machine learning help unleash the alternative food potential of microalgae. Adv Nutr 14(1):1–11. https://doi.org/10.1016/j.advnut.2022.11.002

Heras SC, Jones MN, Gernaey KV, Kruhne U, Mansouri SS (2020) An e-learning bot for bioprocess systems engineering. Comput Aided Chem Eng 48:2023–2028. https://doi.org/10.1016/b978-0-12-823377-1.50338-4

Karim AS, Dudley QM, Juminaga A, Yuan Y, Crowe SA, Heggestad JT, Garg S, Abdalla T, Grubbe WS, Rasor BJ, Coar DN, Torculas M, Krein M, Liew F, Quattlebaum A, Jensen RO, Stuart JA, Simpson SD, Köpke M, Jewett MC (2020) *In vitro* prototyping and rapid optimization of biosynthetic enzymes for cell design. Nat Chem Biol 16(8):912–919. https://doi.org/10.1038/s41589-020-0559-0

Kasemiire A, Avohou HT, De Bleye C, Sacre P-Y, Dumont E, Hubert P, Ziemons E (2021) Design of experiments and design space approaches in the pharmaceutical bioprocess optimization. Eur J Pharm Biopharm 166:144–154. https://doi.org/10.1016/j.ejpb.2021.06.004

Kell DB (2006) Metabolomics, modelling and machine learning in systems biology—towards an understanding of the languages of cells. FEBS J 273(5):873–894. https://doi.org/10.1111/j.1742-4658.2006.05136.x

Khan M, Chuenchart W, Surendra KC, Kumar Khanal S (2023) Applications of artificial intelligence in anaerobic co-digestion: recent advances and prospects. Bioresour Technol 370:128501. https://doi.org/10.1016/j.biortech.2022.128501

Khanal SK, Tarafdar A, You S (2023) Artificial intelligence and machine learning for smart bioprocesses. Bioresour Technol 375:128826. https://doi.org/10.1016/j.biortech.2023.128826

Khuat TT, Bassett R, Otte E, Grevis-James A, Gabrys B (2024) Applications of machine learning in antibody discovery, process development, manufacturing and formulation: current trends, challenges, and opportunities. Comput Chem Eng 182:108585. https://doi.org/10.1016/j.compchemeng.2024.108585

Krishna VVSV, Pappa N, Rani SPJV (2021) Deep learning based soft sensor for bioprocess application. In: 2021 IEEE second international conference on control, measurement and instrumentation (CMI), 2021. https://doi.org/10.1109/cmi50323.2021.9362948

Lancashire LJ, Lemetre C, Ball GR (2008) An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies. Brief Bioinform 10(3):315–329. https://doi.org/10.1093/bib/bbp012

Lee JH, Shin J, Realff MJ (2018) Machine learning: overview of the recent progresses and implications for the process systems engineering field. Comput Chem Eng 114:111–121. https://doi.org/10.1016/j.compchemeng.2017.10.008

Lever J, Krzywinski M, Altman N (2017) Principal component analysis. Nat Methods 14(7):641–642. https://doi.org/10.1038/nmeth.4346

Li Y, Yuan J (2006) Prediction of key state variables using support vector machines in bioprocesses. Chem Eng Technol 29(3):313–319. https://doi.org/10.1002/ceat.200500182

Li W, Gu S, Zhang X, Chen T (2020) Transfer learning for process fault diagnosis: knowledge transfer from simulation to physical processes. Comput Chem Eng 139:106904. https://doi.org/10.1016/j.compchemeng.2020.106904

Liao W, Shen J, Manickam S, Li S, Tao Y, Li D, Liu D, Han Y (2023) Investigation of blueberry juice fermentation by mixed probiotic strains: regression modeling, machine learning optimization and comparison with fermentation by single strain in the phenolic and volatile profiles. Food Chem 405:134982. https://doi.org/10.1016/j.foodchem.2022.134982

Lutnick B, Ramon AJ, Ginley B, Csiszer C, Kim A, Flament I, Damasceno PF, Cornibe J, Parmar C, Standish K, Carrasco-Zevallos O, Yip SSF (2023) Accelerating pharmaceutical R&D with a user-friendly AI system for histopathology image analysis. J Pathol Inform 14:100337. https://doi.org/10.1016/j.jpi.2023.100337

Maharjan R, Hada S, Lee JE, Han H-K, Kim KH, Seo HJ, Foged C, Jeong SH (2023) Comparative study of lipid nanoparticle-based mRNA vaccine bioprocess with machine learning and combinatorial artificial neural network-design of experiment approach. Int J Pharm 640:123012. https://doi.org/10.1016/j.ijpharm.2023.123012

Mahmud M, Kaiser MS, Hussain A, Vassanelli S (2018) Applications of deep learning and reinforcement learning to biological data. IEEE Trans Neural Netw Learn Syst 29(6):2063–2079. https://doi.org/10.1109/tnnls.2018.2790388

Mandenius C-F, Titchener-Hooker NJ (eds) (2013) Measurement, monitoring, modelling and control of bioprocesses. Advances in biochemical engineering/biotechnology. Springer, Berlin. https://doi.org/10.1007/978-3-642-36838-7

Mowbray M, Savage T, Wu C, Song Z, Cho BA, Del Rio-Chanona EA, Zhang D (2021a) Machine learning for biochemical engineering: a review. Biochem Eng J 172:108054. https://doi.org/10.1016/j.bej.2021.108054

Mowbray M, Smith R, Del Rio-Chanona EA, Zhang D (2021b) Using process data to generate an optimal control policy via apprenticeship and reinforcement learning. AIChE J. https://doi.org/10.1002/aic.17306

Mowbray MR, Wu C, Rogers AW, del Rio-Chanona EA, Zhang D (2022) A reinforcement learning-based hybrid modeling framework for bioprocess kinetics identification. Biotechnol Bioeng 120(1):154–168. https://doi.org/10.1002/bit.28262

Nagata Y, Chu KH (2003) Optimization of a fermentation medium using neural networks and genetic algorithms. Biotechnol Lett 25:1837–1842. https://doi.org/10.1023/A:1026225526558

Narayanan H, Dingfelder F, Butté A, Lorenzen N, Sokolov M, Arosio P (2021) Machine learning for biologics: opportunities for protein engineering, developability, and formulation. Trends Pharmacol Sci 42(3):151–165. https://doi.org/10.1016/j.tips.2020.12.004

Natarajan P, Moghadam R, Jagannathan S (2021) Online deep neural network-based feedback control of a Lutein bioprocess. J Process Control 98:41–51. https://doi.org/10.1016/j.jprocont.2020.11.011

Nian R, Liu J, Huang B (2020) A review on reinforcement learning: introduction and applications in industrial process control. Comput Chem Eng 139:106886. https://doi.org/10.1016/j.compchemeng.2020.106886

Nielsen J, Keasling JD (2016) Engineering cellular metabolism. Cell 164(6):1185–1197. https://doi.org/10.1016/j.cell.2016.02.004

Oberhardt MA, Palsson BØ, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol. https://doi.org/10.1038/msb.2009.77

Pan E, Petsagkourakis P, Mowbray M, Zhang D, del Rio-Chanona EA (2021) Constrained model-free reinforcement learning for process optimization. Comput Chem Eng 154:107462. https://doi.org/10.1016/j.compchemeng.2021.107462

Pandey AK, Park J, Ko J, Joo H-H, Raj T, Singh LK, Singh N, Kim S-H (2023) Machine learning in fermentative biohydrogen production: advantages, challenges, and applications. Bioresour Technol 370:128502. https://doi.org/10.1016/j.biortech.2022.128502

Patel N, Corbett B, Trygg J, McCready C, Mhaskar P (2020) Subspace based model identification for an industrial bioreactor: handling infrequent sampling using missing data algorithms. Processes 8(12):1686. https://doi.org/10.3390/pr8121686

Pereira S, Kildegaard HF, Andersen MR (2018) Impact of CHO metabolism on cell growth and protein production: an overview of toxic and inhibiting metabolites and nutrients. Biotechnol J. https://doi.org/10.1002/biot.201700499

Petsagkourakis P, Sandoval IO, Bradford E, Zhang D, del Rio-Chanona EA (2019) Reinforcement learning for batch-to-batch bioprocess optimization. Comput Aided Chem Eng 46:919–924. https://doi.org/10.1016/b978-0-12-818634-3.50154-5

Petsagkourakis P, Sandoval IO, Bradford E, Zhang D, del Rio-Chanona EA (2020) Reinforcement learning for batch bioprocess optimization. Comput Chem Eng 133:106649. https://doi.org/10.1016/j.compchemeng.2019.106649

Petsagkourakis P, Sandoval IO, Bradford E, Galvanin F, Zhang D, del Rio-Chanona EA (2022) Chance constrained policy optimization for process control and optimization. J Process Control 111:35–45. https://doi.org/10.1016/j.jprocont.2022.01.003

Phung R (2019) A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. Appl Sci 9(21):4500. https://doi.org/10.3390/app9214500

Popova M, Isayev O, Tropsha A (2018) Deep reinforcement learning for de novo drug design. Sci Adv. https://doi.org/10.1126/sciadv.aap7885

Rathore AS, Singh SK (2015) Use of multivariate data analysis in bioprocessing. BioPharm Int. https://www.biopharminternational.com/view/use-multivariate-data-analysis-bioprocessing. Accessed 1 Sep 2023

Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149. https://doi.org/10.1109/tpami.2016.2577031

Rivera SL, Karim MN (1992) On-line estimation of bioreactors using recurrent neural networks. IFAC Proc Vol 25(2):159–162. https://doi.org/10.1016/s1474-6670(17)50343-2

Ryu G, Kim GB, Yu T, Lee SY (2023) Deep learning for metabolic pathway design. Metab Eng 80:130–141. https://doi.org/10.1016/j.ymben.2023.09.012

Sampaio PN, Calado CRC (2019) Classification of recombinant *Saccharomyces cerevisiae* cells using PLS-DA modelling based on MIR spectroscopy. In: 2019 IEEE 6th Portuguese meeting on bioengineering (ENBENG), 2019. https://doi.org/10.1109/enbeng.2019.8692463

Sharma V, Tsai M-L, Chen C-W, Sun P-P, Nargotra P, Dong C-D (2023) Advances in machine learning technology for sustainable biofuel production systems in lignocellulosic biorefineries. Sci Total Environ 886:163972. https://doi.org/10.1016/j.scitotenv.2023.163972

Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35(5):1285–1298. https://doi.org/10.1109/tmi.2016.2528162

Smiatek J, Jung A, Bluhmki E (2020) Towards a digital bioprocess replica: computational approaches in biopharmaceutical development and manufacturing. Trends Biotechnol 38(10):1141–1153. https://doi.org/10.1016/j.tibtech.2020.05.008

Takei Y, Ishida T (2022) A benchmark dataset for evaluating practical performance of model quality assessment of homology models. Bioengineering 9(3):118. https://doi.org/10.3390/bioengineering9030118

Tang S, Liao D, Li X, Lin Y, Han S, Zheng S (2021) Cell-free biosynthesis system: methodology and perspective of *in vitro* efficient platform for pyruvate biosynthesis and transformation. ACS Synth Biol 10(10):2417–2433. https://doi.org/10.1021/acssynbio.1c00252

Treloar NJ, Fedorec AJH, Ingalls B, Barnes CP (2020) Deep reinforcement learning for the control of microbial co-cultures in bioreactors. PLoS Comput Biol 16(4):e1007783. https://doi.org/10.1371/journal.pcbi.1007783

Upadhyay A, Upadhyay A, Sarangi PK, Chawade A, Pareek N, Tripathi D, Vivekanand V (2024) Machine learning approach for microbial growth kinetics analysis of acetic acid-producing bacteria isolated from organic waste. Biochem Eng J 202:109164. https://doi.org/10.1016/j.bej.2023.109164

van Heel M (1991) A new family of powerful multivariate statistical sequence analysis techniques. J Mol Biol 220(4):877–887. https://doi.org/10.1016/0022-2836(91)90360-i

van Lent P, Schmitz J, Abeel T (2023) Simulated design–build–test–learn cycles for consistent comparison of machine learning methods in metabolic engineering. ACS Synth Biol 12(9):2588–2599. https://doi.org/10.1021/acssynbio.3c00186

Vega-Ramon F, Zhu X, Savage TR, Petsagkourakis P, Jing K, Zhang D (2021) Kinetic and hybrid modeling for yeast astaxanthin production under uncertainty. Biotechnol Bioeng 118(12):4854–4866. https://doi.org/10.1002/bit.27950

Viana FAC (2015) A tutorial on Latin hypercube design of experiments. Qual Reliab Eng Int 32(5):1975–1985. https://doi.org/10.1002/qre.1924

Wang Z, Peng X, Xia A, Shah AA, Huang Y, Zhu X, Zhu X, Liao Q (2022) The role of machine learning to boost the bioenergy and biofuels conversion. Bioresour Technol 343:126099. https://doi.org/10.1016/j.biortech.2021.126099

Whittle PW (1952) On principal components and least square methods of factor analysis. Scand Actuar J 1952:223–239. https://doi.org/10.1080/03461238.1955.10430696

Widodo A, Yang B-S (2007) Support vector machine in machine condition monitoring and fault diagnosis. Mech Syst Signal Process 21(6):2560–2574. https://doi.org/10.1016/j.ymssp.2006.12.007

Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 58(2):109–130. https://doi.org/10.1016/s0169-7439(01)00155-1

Xiao Z, Li W, Moon H, Roell GW, Chen Y, Tang YJ (2023) Generative Artificial Intelligence GPT-4 accelerates knowledge mining and machine learning for synthetic biology. ACS Synth Biol 12(10):2973–2982. https://doi.org/10.1021/acssynbio.3c00310

Xu P (2019) Analytical solution for a hybrid logistic-monod cell growth model in batch and continuous stirred tank reactor culture. Biotechnol Bioeng 117(3):873–878. https://doi.org/10.1002/bit.27230

Xu P, Dai X, Graf E, Martel R, Russell R (2014) Effects of glutamine and asparagine on recombinant antibody production using CHO-GS cell lines. Biotechnol Prog 30(6):1457–1468. https://doi.org/10.1002/btpr.1957

Yang ZR (2004) Biological applications of support vector machines. Brief Bioinform 5(4):328–338. https://doi.org/10.1093/bib/5.4.328

Yang C-T, Kristiani E, Leong YK, Chang J-S (2023) Big data and machine learning driven bioprocessing—recent trends and critical analysis. Bioresour Technol 372:128625. https://doi.org/10.1016/j.biortech.2023.128625

Zhang Y-HP, Sun J, Ma Y (2017) Biomanufacturing: history and perspective. J Ind Microbiol Biotechnol 44(4–5):773–784. https://doi.org/10.1007/s10295-016-1863-2

Zheng X, Ishikawa J, Sugiyama T, Maruyama Y (2017) Bayesian optimization analysis of containment-venting operation in a boiling water reactor severe accident. Nucl Eng Technol 49(2):434–441. https://doi.org/10.1016/j.net.2016.12.011

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

https://doi.org/10.33472/AFJBS.6.8.2024.200-208

African Journal of Biological Sciences

Research Paper Open Access

Magnesium oxide Nanoparticle synthesis using *Rhizophora lamarckii* plant extract and its characterization

Ankit Kundu[1], Jai Gopal Sharma[2,*]

[1,2]Department of Biotechnology, Delhi Technological University, Delhi-110042, India

Acknowledgement - The authors would like to thank Delhi Technological University for providing infrastructure and facility.

Article History
Volume 6,Issue 8, 2024
Received:20 Feb 2024
Accepted : 28 Mar 2024
doi: 10.33472/AFJBS.6.8.2024.200-208

ABSTRACT

Nanoparticle production is nanotechnology's most important feature. In recent years, physical and chemical techniques to nanoparticle synthesis have shown severe flaws, necessitating the creation of an environmentally benign nanoparticle synthesis methodology. As a result, in this paper, we will detail how to synthesise magnesium oxide nanoparticles (MgONPs) from *Rhizophora lamarckii* and secondary chemicals found in the plant's leaf broth. *Rhizophora lamarckii* is being used to produce MgO nanoparticles by varying parameters such as temperature and pH. A UV-Visible spectrophotometer is used to analyse the absorbance spectra of MgONPs throughout the final synthesis process. SEM, and FTIR and XRD were employed to evaluate the biosynthesized nanoparticles. The current study, which employs a plant to produce MgONPs, is an efficient, environmentally friendly, and straightforward method that could have a significant impact on future industrial and commercial metallic nanoparticle production.

Keywords: Nanoparticles, *Rhizophora lamarckii*, UV-Visible spectrophotometer, Biosynthesized

1. INTRODUCTION

Nanotechnology is playing an increasingly important role in biology, medicine, physics, chemistry, and electronics (1). Agriculture, medicine, waste management, catalysis, pollution control, and therapeutic sectors such as antibacterial and anticancer medicines are all potential applications for nanotechnology. Nanoparticles can have many shapes, sizes,

morphologies, and geometries (2). Physical and chemical approaches to nanoparticle production are damaging to both the environment and humans. The chemical synthesis of nanoparticles has a negative influence on the environment and humans since it emits poisonous hazardous compounds such as sodium borohydride and hydrazine (3). Hydrazine produces irritation in the eyes, throat, and nose. In severe situations, it can cause seizures, coma, and damage to the liver and kidneys (4). Chemical techniques such as microemulsion and evaporation/condensation are equally expensive. Physical methods, such pulse laser ablation, microwave-assisted, and pulsed/explosion wire discharge, need more energy, time, and money (5). As a result, synthesising nanoparticles using green technology appears more practical, non-toxic, cost-effective, and environmentally beneficial (6). It is a bottom-up technique in which plant extract, bacteria, or fungi are employed as raw material and redox reactions occur (7).

Medicinal plants are often used as reducing agents in green nanoparticle synthesis, and they create a wide spectrum of nanoparticles such as copper, silver, gold, zinc, and iron (8). Biologically generated metal nanoparticles have enormous possibilities as cytotoxic agents against cancer cells, antioxidant powder, antibacterial and anticoagulant agents (9). NPs, which are made from a rich pool of secondary metabolites, contain alkaloids and flavonoids. Magnesium nanoparticles offer a wide range of essential uses, particularly in the biological domain, among other metal nanoparticles (10). Magnesium oxide nanoparticles are employed in a variety of applications, including large magnetic resistant materials, sensors, catalysts, optical, electrical, gas sensors, superconductors, solar energy conversion, and the production of organic-inorganic nanostructure composites (11) (12).

The notion of synthesising MgONPs from mangrove plants originated in this setting. Mangrove plants have been used in traditional medicine to cure many ailments. This study suggests a straightforward and efficient methodology for producing MgONPs. Its leaves are useful for treating eye issues and headaches. There is insufficient qualitative and quantitative data to establish the quality and purity of leaves. In this study, magnesium oxide nanoparticles are synthesised using a leaf extract of *Rhizophora lamarckii*. UV-Vis, FTIR, SEM, and XRD techniques are used to characterise the MgONPs that have been synthesised. The plant-mediated synthesis of MgONPs, as mediated by various characterisation approaches, provides substantial insights into the many characteristics that exist today.

MATERIALS AND METHODS

Collection of plant sample

Rhizophora lamarckii leaves were obtained from a forest reserve near Hauz Khas, Delhi. The leaves were thoroughly cleaned with distilled water before being dried on blotting paper. Then, they were sun dried for one day to remove moisture before being pounded into fine powder with a mortar and pestle. This powder was simmered in MilliQ water for 5-10 minutes. Extract obtained after dissolution was used to synthesise MgO nanoparticles.

Preparation of leaf extract

Leaf extract was prepared by weighing 4g of leaf powder and adding it to a reaction beaker containing 100ml MilliQ. The solution was boiled on the heating mantle at 80 degrees for around 10 minutes. After the solution cooled to room temperature, filtration was carried out through Whatman filter paper to produce filtrate. The residue was discarded, and the finished extract was kept at 4°C.

Synthesizing CuO Nanoparticles

All of the chemicals were AR Grade and obtained from reliable providers. MgO nanoparticles were produced using a green technique. Filtered leaf broth was added drop by drop to a 90ml stirring solution containing 10mM magnesium nitrate until the colour changed from yellow to yellowish brown. The liquid was centrifuged at 6000 rpm for 15 minutes. Resulting mixture was left to stir for an hour. It was kept at room temperature overnight and wrapped in aluminium foil before being tested with a range of spectrophotometer techniques.

Characterization

All reagents were acquired from Fisher Scientific Chemicals. The unique properties of nanoparticles were investigated by recording their absorption spectra with an Eppendorf spectrometer. The FTIR spectrum was obtained using a Perkin Elmer instrument and FTIR accessories from PCI analytics, with the synthesised NPs recorded as KBr pellets in the 4000-400 cm^{-1} range (13). The morphological characteristics of the resulting nanoparticles were measured with the SEM ZEISS Instrument EVO 18 Special Edition. SEM images of the synthesised nanoparticles may provide proof of newly generated MgONPs. To assess the stability of synthesised NPs, their size was measured using a Zetasizer and pH was assessed using a HACH device (14). XPERT-PRO X-Ray Diffractometer was used to analyse the results of XRD.

2. RESULTS AND DISCUSSION

Nanoparticles can be created utilising biological agents such as plant extracts, which offers substantial advantages over conventional approaches. It also eliminates the need to monitor cell cultures during large-scale manufacturing. (15). An aqueous leaf extract of *Rhizophora lamarckii* is the most cost-effective method for nucleating nanoparticles. In the current study, MgO NPs were combined with an extricated arrangement of *Rhizophora lamarckii* leaf extract to serve as a dissolvable rather than natural solvent. MgONPs creation can be confirmed using a variety of analytical techniques that provide precise information about nanoparticle synthesis.

UV–Visible Analysis

The nanoparticles were characterised using UV-Vis absorption spectroscopy on an Eppendorf Biospectrometer. Figure 1 depicts the UV-visible spectrum of MgO NPs generated by *Rhizophora lamarckii*. A significant absorption band was detected in the 290-300 nm range, indicating the existence of MgONPs. The absorption spectra alter as the extract composition increases, resulting in production of MgO nanoparticles. As a result, as the content of plant

extract grows, so does the absorption spectrum, with maximum absorption occurring at approximately 300 nm. The observed spectra result from the production of MgO NPs. Even after extended storage, the absorption spectra remained consistent. The collected spectra revealed that the synthesised nanoparticles were symmetrical and spherical in shape.

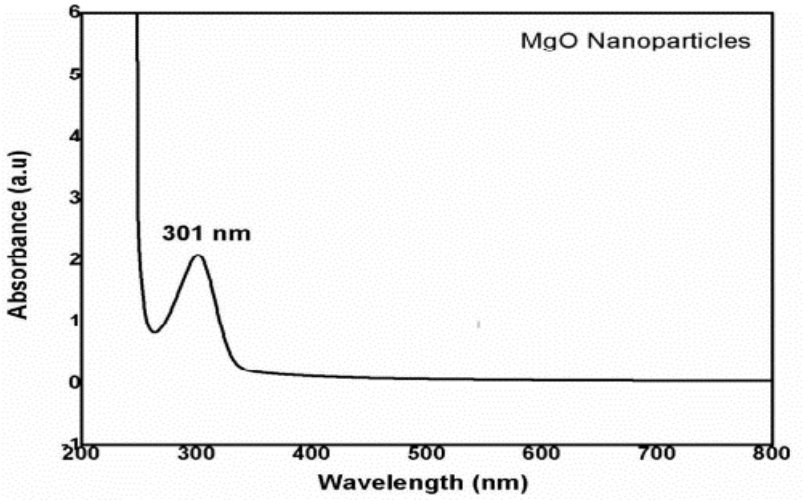


Figure 1: UV-Vis Spectra

Photocatalytic activity

The photocatalytic activity of synthesised MgO nanoparticles was assessed in sunlight using methylene blue dye degradation. Figure 2 displays absorption spectra at different time intervals of irradiation after MgONPs were exposed to sunlight with dye solution and monitored using UV spectra. The greatest absorption spectra in the UV-visible spectrum occur at 660nm. As depicted in the illustration, absorption spectra gradually diminish after irradiation, and the dye disappears within 240 minutes. MgO nanoparticles react with dye to produce electrons and holes, resulting in super oxide radicals. When these hydroxyl radicals reacted with electrons, they destroyed the colour molecules.

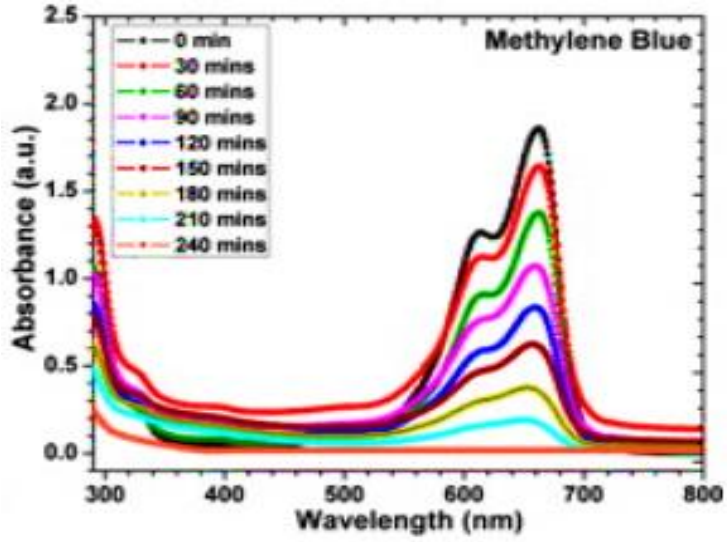


Figure 2: Photocatalytic activity of MgO NPs

XRD Analysis

The resultant nanoparticles were examined with an X-ray diffraction analyzer (Bruker D8, MA, USA). Prior to the XRD investigation, the materials were lyophilized by freezing them in a vial of liquid nitrogen and then subjected to a lyophilizer for 24 hours (FreeZone Freeze dryers, MO, USA) (16). X-ray diffraction could be used to study crystal structure of lyophilized molecules. Figure 3 shows the XRD patterns produced by MgONPs. The XRD analysis may reveal varied diffraction peaks at different intensities for MgO, with the highest at 200. The weak diffraction peaks appeared initially, followed by the greatest peak intensity. The results confirmed previously observed diffraction peaks for MgONPs.

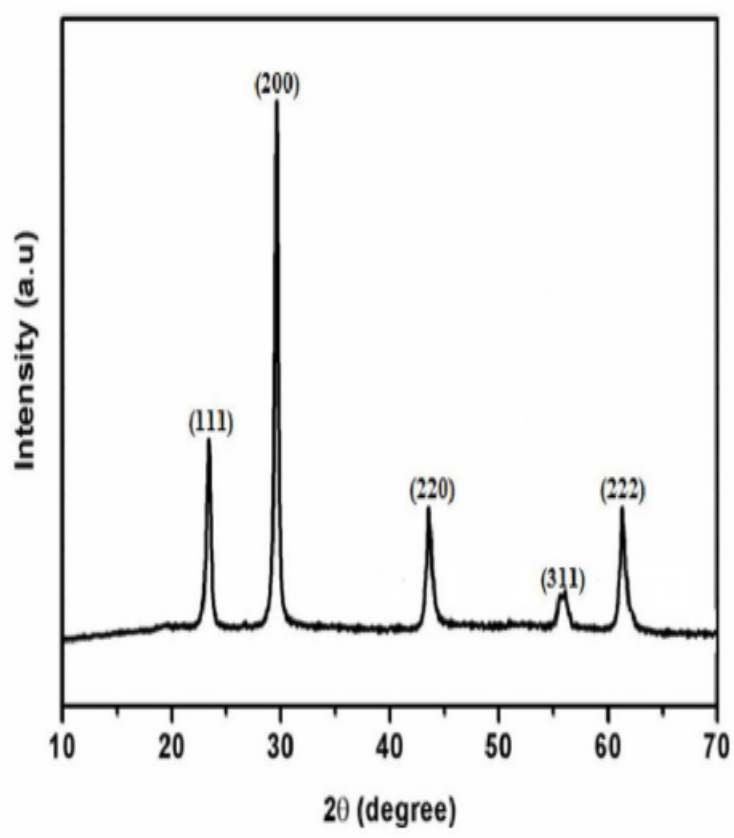


Figure 3: XRD Analysis of MgONPs

FTIR Analysis

FTIR was used to identify biomolecules that decrease and encapsulate MgONPs. Figure 5 shows the FTIR spectra of MgONPs isolated from *R. lamarckii* leaf extract. The spectra have bands at 3405, 2925, 1626, 1402, 1260, and 1035 cm^{-1}, with misplaced peaks at 3395, 2921, 1644, 1450, 1232, and 1077 cm^{-1}. Peaks at 3405 cm^{-1} (shifted to 3395 cm^{-1}) and 1626 cm^{-1} (shifted to 1644 cm^{-1}) show stretching of the N-H amine group. Stretching of the C-H alkane group is shown by absorption maxima shifting from 2925 cm^{-1} to 2921 cm^{-1} and 1402 cm^{-1} to 1450 cm^{-1}. The amine and alkane functional groups of *R. lamarkii* are most likely involved in reducing and capping newly synthesised MgONPs.

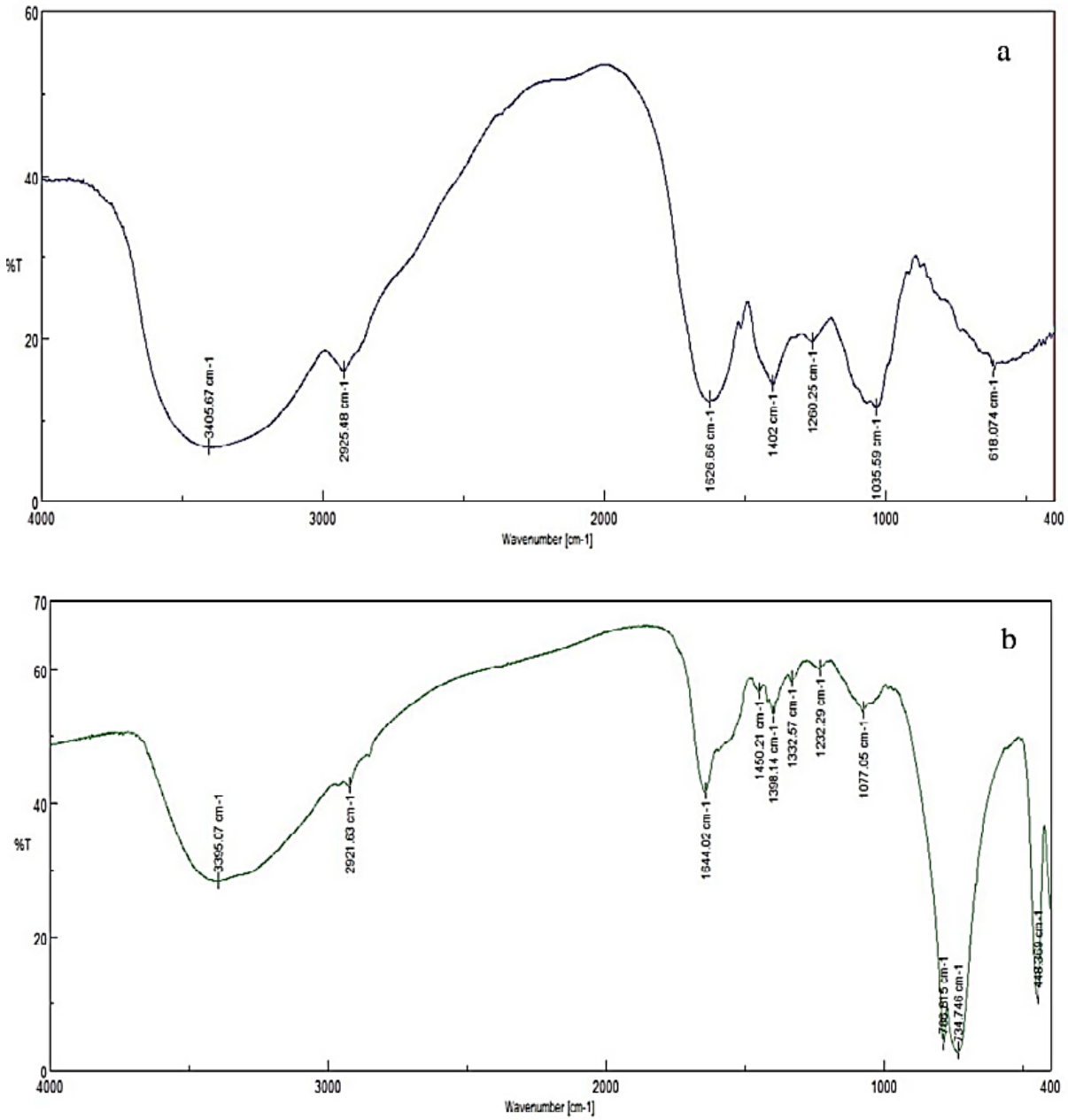


Figure 4: FTIR Analysis of MgONPs: (a) extract of *R. lamarckii* and (b) MgONPs synthesized with *R. lamarckii*

SEM analysis

It is critical to understand the morphological features of the formed nanoparticles by analysing size and shape of the created copper nanoparticles. Scanning electron microscopy was used to analyse the morphological shape and structure of the MgONPs that were created (17). Figure 5 depicts the image produced for SEM investigation of MgONPs. The resultant nanoparticles are generally rod and sphere shaped.

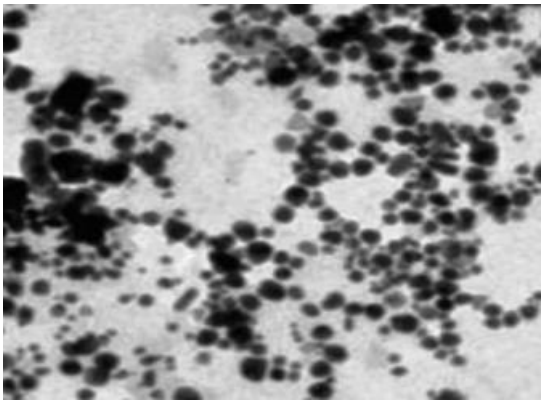


Figure 5: SEM analysis of MgONPs

3. CONCLUSION

Nanoparticle agglomeration can be achieved with a variety of regularly used synthetic substances and techniques. In this study, the amalgamation of copper oxide nanoparticles was explored utilising an environmentally friendly process involving *Rhizophora lamarckii* leaf extracts. Finally, we created a skilled, simple, and successful approach for green amalgamation of MgONPs that employs *Rhizophora lamarckii* leaf extract as a reducing and balancing agent. In compared to the intended MgO NPs, the plant-based MgO nanoparticles were found to be more stable and ejected less particles. Furthermore, whereas hydrodynamic size and accumulation grew over time, plant-blended MgO NPs were smaller and had a lower overall expansion than designed NPs. The proposed MgO NPs were more soluble than plant-interacted MgO NPs, and their particle discharge rate outperformed plant nanoparticles. Plant-incorporated CuO nanoparticles have been demonstrated to be more stable than manufactured nanoparticles. Plant-incorporated MgO NPs were found to be less hazardous than manufactured MgO NPs, making it a better way for releasing more nanoparticles into the environment for a variety of purposes. However, it is uncertain whether the detrimental effects of plant and produced magnesium oxide are caused by particle disintegration or by MgONPs and their interactions with the environment. SEM, XRD, and UV-visible examinations reveal that the MgONPs produced have potential environmental benefits.

Acknowledgment

The authors would like to thank Delhi Technological University for providing infrastructure and facility to carry out the research.

References

1. Roopashree S, Anitha J, Rashmi S. Advent of nanomaterial in modern health science and ayurveda. Mater Today Proc. 2021 Jan 1;46:9096–101.

2. Madkour M, Abdel-Monem YK, Al Sagheer F. Controlled Synthesis of NiO and Co3O4 Nanoparticles from Different Coordinated Precursors: Impact of Precursor's Geometry on the Nanoparticles Characteristics. Ind Eng Chem Res. 2016 Dec 21;55(50):12733–41.

3. León-Silva S, Fernández-Luqueño F, López-Valdez F. Silver Nanoparticles (AgNP) in the Environment: a Review of Potential Risks on Human and Environmental Health. Water Air Soil Pollut. 2016 Aug 5;227(9):306.

4. Tomassoni AJ, French RNE, Walter FG. Toxic Industrial Chemicals and Chemical Weapons. Emerg Med Clin North Am. 2015 Feb;33(1):13–36.

5. Ghorbani HR. Biological and Non-Biological Methods for Fabrication of Copper Nanoparticles. Chem Eng Commun. 2015 Nov 2;202(11):1463–7.

6. Rauwel P, Rauwel E. Emerging Trends in Nanoparticle Synthesis Using Plant Extracts for Biomedical Applications. Glob J Nanomedecin. 2017 Apr 1;1:555562.

7. Jamkhande PG, Ghule NW, Bamer AH, Kalaskar MG. Metal nanoparticles synthesis: An overview on methods of preparation, advantages and disadvantages, and applications. J Drug Deliv Sci Technol. 2019 Oct 1;53:101174.

8. Singh J, Dutta T, Kim KH, Rawat M, Samddar P, Kumar P. 'Green' synthesis of metals and their oxide nanoparticles: applications for environmental remediation. J Nanobiotechnology. 2018 Oct 30;16(1):84.

9. Lateef A, Elegbede J, Akinola O, Ajayi V. Biomedical Applications of Green Synthesized-Metallic Nanoparticles: A Review. 2019 Nov 14;3:157–82.

10. Fahmy H, El-Hakim M, Nady D, Elkaramany Y, Mohamed F, Yasien A, et al. Review on MgO nanoparticles multifunctional role in the biomedical field: Properties and applications. 2022 Jan 1;9:1–14.

11. De Crozals G, Bonnet R, Farre C, Chaix C. Nanoparticles with multiple properties for biomedical applications: A strategic guide. Nano Today. 2016 Aug 1;11(4):435–63.

12. Khan J, Siddiq M, Akram B, Ashraf MA. In-situ synthesis of CuO nanoparticles in P(NIPAM-co-AAA) microgel, structural characterization, catalytic and biological applications. Arab J Chem. 2018 Sep 1;11(6):897–909.

13. Vinosha PA, Mely LA, Jeronsia JE, Krishnan S, Das SJ. Synthesis and properties of spinel ZnFe2O4 nanoparticles by facile co-precipitation route. Optik. 2017 Apr 1;134:99–108.

14. Khan R, Inam MA, Park DR, Zam Zam S, Shin S, Khan S, et al. Influence of Organic Ligands on the Colloidal Stability and Removal of ZnO Nanoparticles from Synthetic Waters by Coagulation. Processes. 2018 Sep;6(9):170.

15. Ghotekar S. A review on plant extract mediated biogenic synthesis of CdO nanoparticles and their recent applications. Asian J Green Chem. 2019 Apr 1;3(2):187–200.

16. Searles JA. Freezing and Annealing Phenomena in Lyophilization. In: Freeze-Drying/Lyophilization of Pharmaceutical and Biological Products. 3rd ed. CRC Press; 2010.

17. Reddy KR. Green synthesis, morphological and optical studies of CuO nanoparticles. J Mol Struct. 2017 Dec 15;1150:553–7.

Journal of Nonlinear Modeling and Analysis
Volume 6, Number 2, June 2024, 485–513
http://jnma-online.com
DOI:10.12150/jnma.2024.485

Meta-Learning-Based Physics-Informed Neural Network: Numerical Simulations of Initial Value Problems of Nonlinear Dynamical Systems Without Labeled Data and Correlation Analyses

Worrawat Duanyai[1], Weon Keun Song[2,†], Thanadol Chitthamlerd[3] and Girish Kumar[4]

Abstract There are several main challenges in solving nonlinear differential equations with artificial neural networks (ANNs), such as a nonlinear system's sensitivity to its initial values, discretization, and strategies for incorporating physics-based information into ANNs. As for the first issue, this paper addresses the initial value problems of nonlinear dynamical systems (a Duffing oscillator and a Burger's equation), which cause large global truncation errors in sub-domains with a significant reduction in the influence of initial constraints, using meta-learning-based physics-informed neural networks (MPINNs). The MPINNs with dual learners outperform physics-informed neural networks with a single learner (no fine reinitialization capability). As a result, the former approach improves solution convergence by 98.83% in the sub-time domain (III) of a Duffing oscillator, and by 85.89% at $t = 45$ in a Burger's equation problem, compared to the latter one. Model accuracy is highly dependent on the adaptability of the initial parameters in the first hidden layers of the meta-models. From correlation analyses, it is obvious that the parameters become less (the Duffing oscillator) or more (the Burger's equation) correlated during fine reinitialization, as the update manner differs or is similar to the one used in pre-initialization. In the first example, the MPINN achieves both the mitigation of model sensitivity to its output and the improvement of model accuracy. Conversely, the second example shows that the proposed approach is not enough to solve both issues simultaneously, as increased model sensitivity to its output leads to higher model accuracy. The application of transfer learning reduces the number of iterative pre-meta-trainings.

[†]the corresponding author.
Email address: worrawatduanyai@gmail.com(W. Duanyai), bauman98@naver.com(W K. Song), thanadolps@gmail.com(T. Chitthamlerd), girish.kumar154@gmail.com(G. Kumar)
[1]Department of Robotic and Computational Intelligence System, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Chalong Krung 1 Rd.,10520 Bangkok, Thailand
[2]Department of Robotic and AI Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Chalong Krung 1, 10520 Bangkok, Thailand
[3]Department of Computer Engineering, Chulalongkorn University, Phaya Thai Rd, Wang Mai, 10330 Bangkok, Thailand
[4]Department of Mechanical Engineering, Delhi Technological University, Shahbad Daulatpur, Bawana Road, 110042 Delhi, India

Keywords Physics-informed neural network, meta-learning, fine reinitialization, transfer learning, initial value problem, nonlinear dynamical systems, correlation analyses

MSC(2010) 65L99, 65M99.

Nomenclature

Symbol	Definition
$\boldsymbol{a}^{[l]}$, $a_n^{[l]}$	Output vector of the l-th hidden layer and the n-th component of $\boldsymbol{a}^{[l]}$
$\boldsymbol{b}^{[l]}$	Bias vector of the l-th hidden layer
c	Constant parameter
cn, dn, sn	Jacobi elliptic functions
d	Constant parameter
$\boldsymbol{d}_k$	Search direction vector at the k-th iteration
$\boldsymbol{D}_{qu}^{(i)}$, $\boldsymbol{D}_{su}^{(i)}$	Query and support datasets, respectively, for the i-th query task and the i-th support one.
$\boldsymbol{D}_{te}, \boldsymbol{D}_{tr}$	Test and training datasets, respectively
f	Constant parameter
$f(\boldsymbol{u})$	Function value at $\boldsymbol{u}$
$f_{\boldsymbol{\theta}}(\boldsymbol{x})$	PINN
$f_{\boldsymbol{\theta}}^m(\boldsymbol{x})$	MPINN
$f_{\boldsymbol{\theta}^{(i)}}^m(\boldsymbol{x}_{su}^{(i)})$	A model for the i-th support task (it is regarded as the i-th support task)
$f_{\boldsymbol{\theta}^{(i)},'}^m(\boldsymbol{x}_{qu}^{(i)})$	A model for the i-th query task (it is regarded as the i-th query task)
$f_{\boldsymbol{\theta}^*}^m(\boldsymbol{x}_{te})$	MPINN evaluated on $\boldsymbol{D}_{te}$
$\boldsymbol{g}_k$	Gradient difference vector at the k-th iteration
$\boldsymbol{H}^{(I)}$, $\boldsymbol{H}_k$	Initial Hessian matrix and Hessian matrix at the k-th iteration
$iter$, $iter^m$	Iterative numbers, respectively, for the PINN (or each inner learner of the MPINN) and the outer learner of the MPINN
$iter^{MAX}$, $iter^{mMAX}$	Maximum iterative numbers, respectively, for the PINN (or each inner learner of the MPINN) and the outer learner of the MPINN.
k	Constant parameter
$L(\boldsymbol{\theta},\ \boldsymbol{D}_{tr})$	Total PINN loss
$L^{(i)}(\boldsymbol{\theta}^{(i)}, \boldsymbol{D}_{su}^{(i)})$, $L^{(i)}(\boldsymbol{\theta}^{(i),'}, \boldsymbol{D}_{qu}^{(i)})$	Support and query losses, respectively, for the i-th task
$\|\|\nabla L(\boldsymbol{\theta}, \boldsymbol{D}_{tr})\|\|$	Norm of the projected gradient of $L(\boldsymbol{\theta}, \boldsymbol{D}_{tr})$
m	Constant parameter
MSE_f, MSE_I, MSE_B, MSE_d	Losses, respectively, of residual, initial constraint, boundary constraint, and labeled data
N_d, N_B, N_f, N_I	Number of labeled data for MSE_d and number of training samples, respectively, for MSE_B, MSE_f and MSE_I
NT	Number of test samples
NQ, NS	Numbers, respectively, of query and support tasks
t, $\triangle t$	Time coordinate and time increment
u, u_n	A single output of $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ and a discrete solution of u at the n-th sampling point
$\boldsymbol{u}$	Output vector of $f_{\boldsymbol{\theta}}(\boldsymbol{x})$
$\bar{u}$, $\bar{u}_n$	Analytical (exact) solution and a discrete solution of $\bar{u}$ at the n-th sampling point
$\boldsymbol{u}_{su}^{(i)}$, $\boldsymbol{u}_{qu}^{(i),'}$, $\boldsymbol{u}_{te}$	Output vectors, respectively, evaluated by $f_{\boldsymbol{\theta}^{(i)}}^m(\boldsymbol{x}_{su}^{(i)})$, $f_{\boldsymbol{\theta}^{(i)},'}^m(\boldsymbol{x}_{qu}^{(i)})$, and $f_{\boldsymbol{\theta}^*}^m(\boldsymbol{x}_{te})$
$\bar{\boldsymbol{u}}_{su}^{(i)}$, $\bar{\boldsymbol{u}}_{qu}^{(i)}$, $\bar{\boldsymbol{u}}_{te}$	Actual output vectors, respectively, for $\boldsymbol{D}_{su}^{(i)}$, $\boldsymbol{D}_{qu}^{(i)}$, and $\boldsymbol{D}_{te}$ ($\bar{\boldsymbol{u}}_{su}^{(i)}$ and $\bar{\boldsymbol{u}}_{qu}^{(i)}$ require no labeled data)
D_{te}^n, ∂_x^n, ∂_t^n	n-th order differential operators
x	Space coordinate
$\boldsymbol{x}$	Input vector

Nomenclature

Symbol	Definition
$\boldsymbol{x}_{su}^{(i)}$, $\boldsymbol{x}_{qu}^{(i)}$, $\boldsymbol{x}_{te}$	Input vectors, respectively, for $\boldsymbol{D}_{su}^{(i)}$, $\boldsymbol{D}_{qu}^{(i)}$, and $\boldsymbol{D}_{te}$
$\boldsymbol{z}^{[l]}$, $z_n^{[l]}$	Output vector of the l-th hidden layer before taking an activation and the n-th component of $\boldsymbol{z}^{[l]}$.
α	Scalar
α_k	Iterative step size
β	Learnable parameter for swish
ε, ε^m, ε^H	Convergence tolerances, respectively, for the PINN (or each inner learner of the MPINN), the outer learner of the MPINN, and loss reduction tolerance
$\boldsymbol{\theta}$	Parameter matrix of the PINN and MPINN
$\boldsymbol{\theta}_k$, $\boldsymbol{\theta}_{mk}^{(i),\prime}$	Parameter matrices updated, respectively, at the k-th iteration by the PINN (or each inner learner of the MPINN) and at the mk-th meta-iteration by the outer learner of the MPINN
$\boldsymbol{\theta}^{(I)}$	Initial parameter matrix
$\boldsymbol{\theta}^{(i)}$	Parameter matrix trained on $\boldsymbol{D}_{su}^{(i)}$
$\boldsymbol{\theta}^{(i),\prime}$	Parameter matrix trained on $\boldsymbol{D}_{qu}^{(i)}$
$\boldsymbol{\theta}^{[l-1]}$	Parameter matrix that maps from $\boldsymbol{a}^{[l-1]}$ to $\boldsymbol{z}^{[l]}$
$\theta_{j,i}^{[l-1]}$	Parameter that controls a function mapping from $a_i^{[l-1]}$ to $z_j^{[l]}$
$\boldsymbol{\theta}^*$	Optimal parameter matrix achieved by the meta-training for a meta-test phase
λ_f, λ_I, λ_B, λ_d	Parameters, respectively, for MSE_f, MSE_I, MSE_B, and MSE_d
ω	Constant parameter
μ	Constant parameter
$\boldsymbol{\sigma}^{[l]}$	Activation of the l-th hidden layer

1. Introduction

Physical phenomena of dynamical systems are modeled by governing differential equations, and the analysis of their solutions provides insight into them. That is why it is critical to find precise solutions in science and engineering. Due to the absence of analytical solutions in many practical cases, numerical approaches such as finite difference, finite volume, and finite element methods have dominated over the last seven decades. Nonetheless, classical solvers have disadvantages that require a lot of computational efforts, especially for multi-scale or multi-physics-based nonlinear systems. Researchers continue to face challenges in numerical solver design with the balance between accuracy and robustness. Recently, various artificial neural network (ANN)-based solvers among many alternatives have gained large attention because of the prospect of replacing classical solvers. The advantages of ANNs for solving governing differential equations are as follows: (i) powerful modeling for interdisciplinary problems; (ii) simple modeling for strong nonlinear relationships; (iii) efficient forward evaluation of a trained model for real-time applications; (iv) automatic gradient-based optimization due to analytically differentiable modeling; (v) meshless modeling for discretization. ANNs have been applied to solve dynamical systems governed by not only ordinary differential equations (ODEs) [1–5] but partial differential equations (PDEs) [2, 6–19]. Consistent challenges [20–22] were reported for performance improvement during 1990 – 2000, but there was no significant progress. Raissi et al. [7] proposed a new conceptual neural network known as a physics-informed neural network (PINN) in 2019. They developed a physics-informed fully connected neural network (PIFCNN) with high

model accuracy, where a PDE residual loss was incorporated into data-driven functional representation as one of constraint conditions, in small data regimes. Their success in the research has led many researchers to explore PINNs. There are several types of neural networks in modeling PINNs, and these two are among them: a PIFCNN [2, 3, 5–11, 13, 15–17, 19, 23] and a physics-informed convolutional neural network (PICNN) [4, 8, 12, 14]. Several studies have revealed that PIFCNNs perform well with small data [4, 7, 11] or even in the absence of data [8, 9, 12]. A literature review by us supports the claim that PINNs have addressed various problems in engineering, science, and mathematics: structural dynamics [1], nonlinear dynamical systems in applied mathematics [2, 3, 6, 10, 12–14, 23], seismic response [4], quantum mechanics, reaction-diffusion, and propagation of water wave [7], flow [7–9, 11, 16], thermal-transfer related problems [15, 18], composite material [17]. Studies on differential equations have been mostly engaged in boundary value problems (BVPs) [2, 8, 13–15] and initial boundary value problems (IBVPs) [2, 3, 6, 7, 9–12, 16–19]. When selecting a research subject, the complexity of boundary conditions can motivate researchers to choose it as a highly challenging topic. However, in many cases, solution convergence with some degree of accuracy or even higher accuracy can be achieved by adding intricate boundary constraints without labeled data. While initial value problems (IVPs) [3, 5, 23] have received less attention despite their poor solution convergence in sub-domains with a sharp decline in the influence of initial constraints. To address this fundamental issue, several studies introduced noticeable but limited approaches. Meng et al. [3] proposed a parallel PINN to decompose a long-time domain into multiple independent sub-domains with shorter time interval. This strategy resulted in faster convergence and reasonable solution predictions. However, the optimal number of sub-domains to be applied, depending on the characteristics of dynamic systems, was not discussed. The study by Florio et al. [5] highlighted the combination of a shallow PINN and the theory of functional connections and an extreme learning process. They succeeded in solving the stiff systems of ODEs by the proposed approach without reducing the stiffness of the problems. Nonetheless, the use of labeled data was exposed as a crucial limitation because of expensive data generation. Robinson et al. [23] used the injection of partially known physical information at an intermediate layer in a PINN during training for solving benchmark problems in the literature including oscillations. Despite the focus, the relationship between the injection and the nature of models has remained inconclusive. This implies that discovering a universal method for diverse dynamical systems is a huge challenging task. This study represents the first application of both meta-learning and transfer learning, which involve the fine initialization of parameters, to PINNs for solving the IVPs of a nonlinear Duffing oscillator and a nonlinear Burger's equation without labeled data. A Duffing oscillator is a type of nonlinear second-order ODE that describes the behavior of a damped-driven harmonic oscillator. It is a type of dynamical system that exhibits complex behavior, including chaos, resonance, and bifurcations due to its nonlinear nature. Applications include electronic circuits, mechanical systems, and biological systems, while Burgers' equation is a fundamental nonlinear PDE that governs fluid flow. It is used in various areas of fluid dynamics, including the study of turbulence, shock waves, and nonlinear wave propagation. Additionally, it has applications in other fields, such as traffic flow modeling and nonlinear acoustics. In a broad sense, meta-learning may be extended to include the selection of a suitable algorithm by taking into consideration the features of

input datasets, exploring the topology of neural network architectures, and tuning a set of hyperparameters. In a narrow sense, it is the algorithm to fine-tune a set of model parameters. Our current study is limited to the latter, but we will extend it to the former. A meta-learning strategy for a new task was proposed by Peng et al. [24] who were one of the pioneers to discover a suitable learning algorithm. A typical strategy involves training existing models on a new target dataset and then selecting the model that performs best on it. Such a strategy needs big datasets to achieve adequate model accuracy. However, in many real circumstances, a lack of data would limit the ineffective traditional method to laboratory experiments. A new requirement in machine learning is to discover new algorithms to imitate how humans learn new tasks based on their prior knowledge. As humans learn new similar tasks based on it, a set of fine meta-tuned initial parameters is similar in function to it. This means that meta-learning-based algorithms can learn new analogous features rapidly and precisely with fewer data. Therefore, meta-learning is one of the strategies for how an ANN acquires useful prior knowledge by adjusting initial parameters. Learning rules of meta-learning were proposed by a series of studies in 1991 [25] and 1997 [26]. Hochreiter et al. [27] and Younger et al. [28] marked a turning point in gradient descent-based deep meta-learning that belongs to one of the three categories: metric-based, model-based, and optimization (gradient)-based techniques. Many trials by the third method update a meta-model using a bi-level optimization process with multi-inner learners for tuning support models and a single outer learner for meta-learning. Model-agnostic meta-learning (MAML), proposed by Finn et al. [29], showed the feasibility of an application of the dual learners to few-shot problems for classification, regression, and reinforcement learning using gradient-based backpropagation. Their application was successful in adapting a meta-model to new tasks. Several ideas [30–33] on meta-learning have been proposed to improve the training efficiency of PINNs. This paper discusses the model accuracy of classical PINNs and meta-learning-based PINNs (MPINNs) for the IVPs of both examples. Non-uniform random sampling (NURS) is applied to both in training. Evaluations on the convergence of the solutions to the problems are made across entire domains. Experimental data show that the MPINNs outperform the PINNs in tracking their solutions in all domains as well as in the sub-time domain (III) of the Duffing oscillator and at $t = 45$ in the Burger's equation problem. The meta-learning with the dual learners makes a large contribution to improving the solution convergence of the MPINNs, as it mostly updates the initial values of the first hidden layers of the meta-models. Correlation analyses assist us in gaining a deeper comprehension of how the initial parameters become adaptable to inputs and responsive to outputs throughout the meta-training, depending on the nature of the dynamical systems. The Duffing oscillator achieves both the mitigation of the MPINN's sensitivity to its output and the enhancement of the model accuracy, but regarding the Burger's equation, the proposed model is insufficient to address both issues simultaneously, as the model's sensitivity to its output increases during the meta-tuning. This means that it could be less robust. In the MPINN algorithm, the application of transfer learning to support tasks minimizes the number of iterative pre-meta-tunings.

2. Problem statement and examples

Using PINNs for solving IVPs may result in significant truncation errors due to their limited capacity to accurately adjust initial parameters. Figure 3 and 4 show two of the cases. This paper employs the MPINNs to tackle the issue at hand for two case studies regarding the nonlinear Duffing oscillator and the nonlinear Burger's equation. The IVPs and the BVPs (or the IBVPs) are both treated to examine the characteristics of the examples. When dealing with IVPs, the complexity of initial conditions can cause an initial constraint effect to persist throughout an entire domain. This can lead to insufficient global truncation error occurrence and poor observation of the reduction of the error. Additionally, as neural network modeling becomes more intricate, there can be more factors that influence model accuracy. To avoid these potential factors and any ambiguity in identifying the sources of errors, this study employs the simple examples and concentrates on the analysis of the impact of initial parameters on the solution accuracy of the MPINNs. In addition, it focuses on the correlation analysis between them before and after the fine reinitialization. For the IVPs, the initial condition of the Duffing oscillator is two-point, whereas in the Burger's equation case, it is one-dimensional. Duffing oscillators governed by a second-order nonlinear ODE can capture the relationship between the input and the output of nonlinear dynamical systems under various conditions such as steady-state, transient, and other dynamic scenarios. The behaviour of the exact solution to Eq.(2.1), including amplitude and phase, depends on the parameter value, $1/4$ of $cn(2t, 1/4)$. As it deviates from zero, the function deviates from a simple cosine wave, and the shape and characteristics of the periodic oscillation change accordingly. In other words, the modulus can range from 0 to 1, with values being close to 0 corresponding to simple cosine-like oscillations, and values being close to 1 corresponding to more complex and highly nonlinear oscillations. Figure 3 demonstrates the phenomena mentioned above well.

$$\begin{aligned} &m \cdot D_t^2 \bar{u} + c \cdot D_t \bar{u} + k \cdot \bar{u} + f \cdot \bar{u}^3 = d \cdot \cos(\omega t),\ \forall t \in [0, 9] \\ &\text{with } \bar{u}(0) = 1 \text{ and } D_t \bar{u}(0) = 0. \end{aligned} \tag{2.1}$$

A Burger's equation (see Eq.(2.2)) is a nonlinear PDE governing the behaviour of traffic, liquid, and gas flow as well as the flow of materials during material processing operations such as casting and forging. It has numerous applications in physics, engineering, and applied mathematics.

$$\begin{aligned} &\partial_t \bar{u} + \bar{u} \cdot \partial_x \bar{u} = \mu \cdot \partial_x^2 \bar{u},\ \forall (t, x) \in [0, 90] \times [-4, 4] \\ &\text{with } \bar{u}(0, x) = 2x, \end{aligned} \tag{2.2}$$

here, $m = 1$, $c = 0$, $k = f = 2$, $d = 0$, $\omega = 0$ and $\mu = 1$. Eq.(2.1) has the exact solution of $\bar{u}(t) = cn(2t, \frac{1}{4})$. The exact solution to Eq.(2.2) is $\bar{u}(t, x) = \frac{2x}{1+2t}$. For the BVP of the Duffing oscillator, the boundary conditions are enforced to Eq.(2.1) instead of the given initial conditions: $\bar{u}(0) = 1$ and $D_t \bar{u}(0) = 0$, and $\bar{u}(9) = cn(18, \frac{1}{4})$ and $D_t \bar{u}(9) = -2dn(18, \frac{1}{4})\ sn(18, \frac{1}{4})$. For the IBVP of the Burger's equation, the boundary condition of $\bar{u}(90, x) = \frac{2x}{181}$ is enforced additionally to Eq.(2.2).

3. PINN

This section introduces a model representation of a classical PINN (see Figure 1) and discusses the outputs from the IVPs and the BVP (or the IBVP) including the limitation in applying the PINN model to the IVPs.

3.1. PINN representation

A PINN is a form of multi-layer perceptron that is interpreted as a functional approximator. It can be defined by

$$\bar{\boldsymbol{u}} = f_\theta(\boldsymbol{x}) \tag{3.1}$$

for the Duffing oscillator,

$$\boldsymbol{x} = \{t\} \text{ and } \boldsymbol{u} = \{u\} \tag{3.2}$$

and in the case of the Burger's equation,

$$\boldsymbol{x} = \begin{Bmatrix} t \\ x \end{Bmatrix} \text{ and } \bar{\boldsymbol{u}} = \{u\}. \tag{3.3}$$

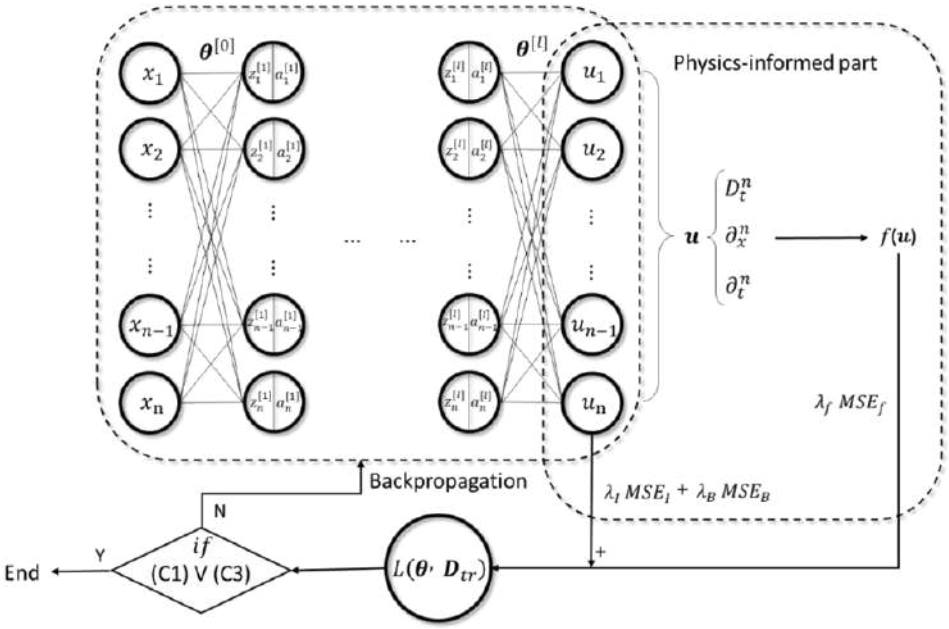


Figure 1. A schematic algorithm of a classical PINN.

An input vector for the first hidden layer is $\boldsymbol{x}$ (see Eq.(3.2) and(3.3)) and each hidden layer's output is represented in a forward propagation phase by

$$\begin{aligned}
\boldsymbol{a}^{[0]} &= \boldsymbol{x}, \\
\boldsymbol{z}^{[1]} &= \boldsymbol{\theta}^{[0]} \cdot \boldsymbol{a}^{[0]} + \boldsymbol{b}^{[1]}, \\
\boldsymbol{a}^{[1]} &= \boldsymbol{\sigma}^{[1]}(\boldsymbol{z}^{[1]}), \\
&\vdots \\
\boldsymbol{z}^{[l]} &= \boldsymbol{\theta}^{[l-1]} \cdot \boldsymbol{a}^{[l-1]} + \boldsymbol{b}^{[l]}, \\
\boldsymbol{a}^{[l]} &= \boldsymbol{\sigma}^{[l]}(\boldsymbol{z}^{[l]}), \\
\boldsymbol{u} &= \boldsymbol{\theta}^{[l]} \cdot \boldsymbol{a}^{[l]} + \boldsymbol{b}^{[l+1]}.
\end{aligned} \tag{3.4}$$

The PINN has a single input layer, a single output one, and l hidden ones. In this study, a Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm is used for optimization. It is in the family of Quasi-Newton methods that update an approximate $\boldsymbol{H}_k$ based on $\boldsymbol{g}_k$ across iterations. The optimizer has three different stopping criteria which are listed below, and it stops when one of those is met: (C1) $iter = iter^{MAX}$; (C2) $L(\boldsymbol{\theta}, \boldsymbol{D}_{tr}) < \varepsilon$; (C3) $||\nabla L(\boldsymbol{\theta}, \boldsymbol{D}_{tr})|| < \varepsilon^H$. Here, $iter$, $iter^{MAX}$ and ε are replaced, respectively, with $iter^m$, $iter^{mMax}$, and $\varepsilon^m$for the meta-training. $L(\boldsymbol{\theta}, \boldsymbol{D}_{tr})$ is also switched with $L^{(i)}(\boldsymbol{\theta}^{(i)}, \boldsymbol{D}_{su}^{(i)})$ and $\sum_{i=1}^{NQ} L^{(i)}(\boldsymbol{\theta}^{(i),'}, \boldsymbol{D}_{qu}^{(i)})$, respectively, for each inner learner and the outer learner of the MPINN. The losses are defined by the same combination of the individual losses of Eq.(3.5). (C2) is applied only to the outer learner, not to each inner learner and the PINN with a single learner. The simulations in the study use the options: $iter^{mMax}$ is 5,000; $iter^{mMax}$ is 50; ε^m is 8.0×10^{-6}; ε^H is 2.22×10^{-9}; the maximum number of line search steps per iteration is 50. The PINN updates $\boldsymbol{\theta}$ by minimizing Eq.(3.5) that includes four individual losses presented below. $\bar{u}_s$ can be acquired through measurements or physical laws. In this research, the MSE_d is not considered.

$$L(\boldsymbol{\theta}, \boldsymbol{D}_{tr}) = \lambda_f MSE_f + \lambda_I MSE_I + \lambda_B MSE_B + \lambda_d MSE_d, \tag{3.5}$$

where $\lambda_f = \lambda_I = \lambda_B = \lambda_d = 1$ and

$$MSE_f = \frac{1}{N_f} \sum_{p=1}^{N_f} [f(u_p)]^2, \tag{3.6a}$$

$$MSE_I = \frac{1}{N_I} \sum_{q=1}^{N_I} [u_q - \bar{u}_q]^2, \tag{3.6b}$$

$$MSE_B = \frac{1}{N_B} \sum_{r=1}^{N_B} [u_r - \bar{u}_r]^2, \tag{3.6c}$$

$$MSE_d = \frac{1}{N_d} \sum_{s=1}^{N_d} [u_s - \bar{u}_s]^2. \tag{3.6d}$$

3.2. Experiments for the performance evaluation of the PINN

In this section, the PINNs with different architectures are used to solve the IVPs and the BVP (or the IBVP), and the examination of the convergence of their solutions is included. The PINN for modelling the Duffing oscillator has six hidden layers, and each has 64 nodes with batch normalization and dropout (dropout rate = 0.5). Initialization is achieved by a variance scaling (VS) initializer and activations are swish with $\beta = 1.0$. For the Burger's equation, it has five hidden layers with 128 nodes per layer. Batch normalization is applied to every single hidden layer. A combination of a normal Xavier initializer and tanh activations is used. Figure 2 and tables 1–2 illustrate the evidence of the use of the NURS to the sub-domains across the entire domains in training the PINNs. The entire domain of the Duffing oscillator is divided into three sub-domains (see Figure 2(a) and Figure 2(b)), while the one of the Burger's equation has four sub-domains (see Figure 2(c) and 2(d)). It functions well on imposing samples in a more concentrated manner neighboring the left end or both ones of the entire domains depending on the nature of the problems. Regarding the Duffing oscillator, the IVP uses only two samples to the MSE_I. Each of (I), (II), and (III) has 84.99%, 10%, and 5% of 7,001 samples, respectively, for the MSE_f. The BVP uses each two samples at t=0 and t=9 for the MSE_B, and each sub-domain uses 42.83%, 14.28%, and 42.83% of 7,004 ones, respectively, for the MSE_f. In the case of the Burger's equation, the MSE_I uses 10,000 samples for the IVP. Each of (I), (II), (III), and (IV) has 38.82%, 1.18%, 0.59%, and 0.59% of 17,000 samples, respectively, for the MSE_f. Whereas the MSE_I only uses 1,000 samples for the IBVP, and it must be much less in comparison to the 10,000 samples of the IVP case. Each of the sub-domains has 81.20%, 1.25%, 1.25%, and 3.75% of 8,005 samples, respectively, for the MSE_f. Each of the intersections (90, -4), (90, -2), (90, 0), (90, 2), and (90, 4) imposed along the right end boundary has a single sample for the MSE_B. It turns out that the IBVP needs much fewer samples than 17,000 ones of the IVP for the highly accurate solution. The Duffing oscillator is tested using 1,000 uniformly distributed samples throughout the domain, and the Burger's equation is sampled specifically along the dashed lines shown in (c) or (d) of Figure 2.

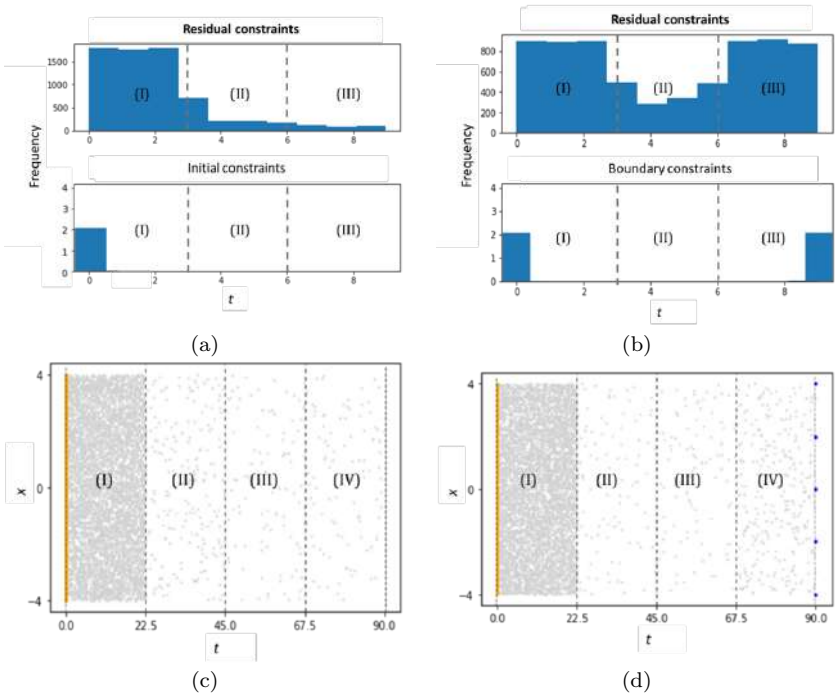


Figure 2. Distributions of samples across the sub-domains. (a) and (b) are, respectively, for the IVP and BVP of the Duffing oscillator; (c) and (d) are, respectively, for the IVP and IBVP of the Burger's equation. In Figure (c) and (d), orange and gray indicate samples, respectively, for MSE_I and MSE_f, and blue displays samples for MSE_B.

Table 1. Number of samples for the Duffing oscillator

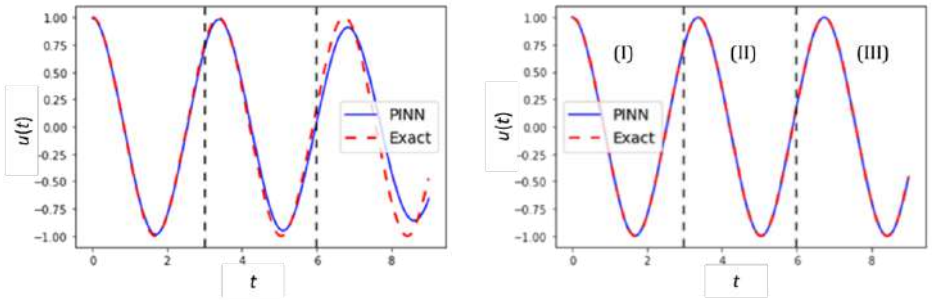
Problem type	MSE_I	MSE_f				MSE_B	Total
		(I)	(II)	(III)	Sub-total		
IVP	2	5,950	700	350	7,000	–	7,001
BVP	–	3,000	1,000	3,000	7,000	4	7,004

Table 2. Number of samples for the Burger's equation

Problem type	MSE_I	MSE_f					MSE_B	Total
		(I)	(II)	(III)	(IV)	Sub-total		
IVP	10,000	6,600	200	100	100	7,000	–	17,000
BVP	1,000	6,500	100	100	300	7,000	5	8,005

To discuss the solution accuracy of the MPINNs as well as the PINNs, we quantify the global truncation errors using the root mean square error (RMSE) presented below:

$$RMSE = \sqrt{\frac{1}{NT}\sum_{n=1}^{NT}(u_n - \bar{u}_n)^2}. \tag{3.7}$$



(a) (b)

Figure 3. Convergence of the solutions to the Duffing oscillator. (a) Comparison of the solution by the PINN and the exact one to the IVP. (b) Comparison of the solution by the PINN and the exact one to the BVP.

Table 3. Comparison of the RMSEs of the Duffing oscillator

Problem	RMSE(m)			
type	(I)	(II)	(III)	Overall
IVP	1.673×10^{-2}	5.967×10^{-2}	1.697×10^{-1}	1.043×10^{-1}
BVP	7.443×10^{-4}	1.307×10^{-3}	6.192×10^{-4}	9.391×10^{-4}

Figure 3 and Table 3 indicate that the RMSEs of the IVP increase with t, and (III) is the critical sub-domain. Unlike the IVP, the largest RSME occurs in (II) for the BVP, but it is relatively very small. The overall RMSE of the IVP increase by 11,006.4 %, and the local RMSE in (III) rises by 27,306.3 % in comparison to the BVP. As shown in Figure 4 and Table 4, the sub-domains at $t = 45$ and $t = 67.5$ are critical for the IVP of the Burger's equation. The IBVP finds the critical sub-domain in the middle, but it is also relatively small. The overall RMSE of the IVP increases by 659.69 %, and the local RMSE at $t = 45$ rises by 475.87 % compared to the IBVP. When it comes to the IVPs, it is obvious from the findings that the PINNs tend to increase the local RMSEs over t. This trend is more pronounced in the case of the Duffing oscillator. In the case of the Burger's equation, the local RMSEs start to increase rapidly after $t = 45$ and decrease somewhat around $t = 90$, but the decrement is not significant. This phenomenon appears to be mainly caused by a reduction in the influence of the initial constraints. In contrast, the boundary constraints play an imperative role in the high-quality convergence of the solutions for the BVP and the IBVP, especially in the critical sub-domains. One of the driving forces behind this study is the IVP issue. In section 4, we propose an alternative strategy to address it for the examples.

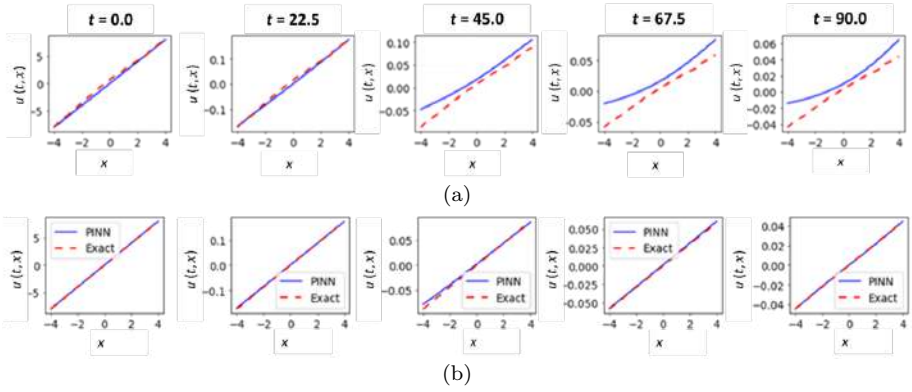


Figure 4. Convergence of the solutions to the Burger's equation. (a) Comparison of the solution by the PINN and the exact one to the IVP. (b) Comparison of the solution by the PINN and the exact one to the IBVP.

Table 4. Comparison of the RMSEs of the Burger's equation

Problem	RMSE(m/s)					
type	$t = 0$	$t = 22.5$	$t = 45$	$t = 67.5$	$t = 90$	Overall
IVP	1.057×10^{-3}	2.198×10^{-3}	2.277×10^{-2}	2.247×10^{-2}	1.616×10^{-2}	1.293×10^{-2}
IBVP	7.837×10^{-4}	2.101×10^{-3}	3.954×10^{-3}	1.271×10^{-3}	3.973×10^{-4}	1.702×10^{-3}

4. MPINN

In machine learning, there are several strategies for better adaptation of parameters to new inputs by learning from prior models. In this regard, various meta-learning techniques have been introduced in the literature [34]. This section discusses the MPINN based on the MAML that fine-reinitializes $\boldsymbol{\theta}$ using bi-level optimization. Comparative evaluations on the convergence of the solutions to the IVPs by the PINNs and MPINNs are implemented.

4.1. Meta-learning based on the MAML

An ANN is a typical nonlinear system, and its solution is highly dependent on its initial parameter values, as is well known. A meta-learning process is designed to improve the adaptability of the initial values of $\boldsymbol{\theta}$ that leads to better solution convergence. To achieve this, bi-level optimization is required. The outer learner, which reinitializes a meta-model for each inner learner in every single meta-iteration, enables each inner one to learn more adaptable features to a new task than a single learner. As a result, bi-level optimization can improve model accuracy as $\boldsymbol{\theta}$ alters its direction and magnitude of increments in a different update manner (see Figure 5), but to improve it, $\boldsymbol{\theta}$ can also be updated in a similar manner, as discussed in section 5 for the Burger's equation. Figure 5 summarizes (a) of Figure 6. Figure 5 and 6 display that first, training on $\boldsymbol{D}_{su}^{(i)}$ updates $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^{(i)}$ by adapting $\boldsymbol{\theta}$ to $\boldsymbol{x}_{su}^{(i)}$. $\boldsymbol{\theta}^{(i)}$ takes a next forward step and achieves $\boldsymbol{\theta}^{(i),'}$ when $\sum_{i=1}^{NQ} L^{(i)}(\boldsymbol{\theta}^{(i),'}, \boldsymbol{D}_{qu}^{(i)})$ minimizes. Therefore, $\boldsymbol{\theta}$ reaches $\boldsymbol{\theta}^{(i),'}$ through $\boldsymbol{\theta}^{(i)}$. $\boldsymbol{\theta}^*$ is more adjustable to a new task when

$\boldsymbol{D}_{su}^{(i)}$ and $\boldsymbol{D}_{qu}^{(i)}$ are more similar in characteristics to $\boldsymbol{D}_{te}$.

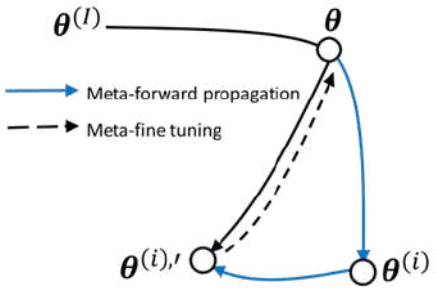


Figure 5. The key concept of the MAML for a single new task.

In the meta-training, the BFGS optimizer is also used. The first step begins with both $\boldsymbol{\theta}^{(I)}$ and $\boldsymbol{H}^{(I)}$, and a procedure with the steps of 1 - 6 for each of $f^m_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{x}^{(i)}_{su})$ is repeated until one of the criteria stops the optimizer. The BFGS keeps $\boldsymbol{H}_k$ positive definite for updating formulations:

$$\text{Step 1: find } \boldsymbol{d}_k \text{ by solving } \boldsymbol{H}_k\boldsymbol{d}_k = -\nabla L(\boldsymbol{\theta}_k, \boldsymbol{D}_{tr}). \tag{4.1a}$$

$$\text{Step 2: perform a line search in } \boldsymbol{d}_k \text{ for computing } \alpha_k = \alpha \text{ by minimize } L(\boldsymbol{\theta}_k + \alpha\boldsymbol{d}_k, \boldsymbol{D}_{tr}). \tag{4.1b}$$

$$\text{Step 3: set } \boldsymbol{s}_k = \alpha_k\boldsymbol{d}_k \text{ and update } \boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \boldsymbol{s}_k. \tag{4.1c}$$

$$\text{Step 4: compute } \boldsymbol{g}_k = \nabla L(\boldsymbol{\theta}_{k+1}, \boldsymbol{D}_{tr}) - \nabla L(\boldsymbol{\theta}_k, \boldsymbol{D}_{tr}). \tag{4.1d}$$

$$\text{Step 5: update } \boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\boldsymbol{g}_k\boldsymbol{g}_k{}^T}{\boldsymbol{g}_k{}^T\boldsymbol{s}_k} - \frac{\boldsymbol{H}_k\boldsymbol{s}_k\boldsymbol{s}_k{}^T\boldsymbol{H}_k{}^T}{\boldsymbol{s}_k{}^T\boldsymbol{H}_k\boldsymbol{s}_k}. \tag{4.1e}$$

Step 6: return to step 1.

In (a) of Figure 6, $f^m_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{x}^{(i)}_{su})$ is deployable and NS is nine for both examples, and NQ becomes automatically equal to NS. $\boldsymbol{\theta}$ is updated to $\boldsymbol{\theta}^{(i)}$ by each inner learner when the update of $\boldsymbol{\theta}_k$ in Eq.(4.1c) on $\boldsymbol{D}^{(i)}_{su}$ is completed for $f^m_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{x}^{(i)}_{su})$. Each inner learner follows the learning rule of the PINN. The outer learner updates $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^{(i),\prime}$ ($\boldsymbol{\theta}^{(i),\prime}$ means $\boldsymbol{\theta}^*$ in the test phase) in substance through $\boldsymbol{\theta}^{(i)}$. It is also achieved by Eq.(4.1c) when Eq.(4.2) instead of $L(\boldsymbol{\theta}_k, \boldsymbol{D}_{tr})$ is applied to Eq.(4.1a). Eq.(4.2) indicates the total of individual query losses.

$$\sum_{i=1}^{NQ} L^{(i)}(\boldsymbol{\theta}^{(i),\prime}_{mk}, \boldsymbol{D}^{(i)}_{qu}). \tag{4.2}$$

The optimizer stops the outer learner by (C3) when each $iter^m$ reaches 17 and 40, respectively, for the Duffing oscillator and the Burger's equation. The transfer of the parameters (see (b) of Figure 6) significantly drops the number of the iterative pre-meta-tunings. For the Duffing oscillator, $f^m_{\boldsymbol{\theta}^{(1)}}(\boldsymbol{x}^{(1)}_{su})$ undergoes 1,077 iterations for the training, but $f^m_{\boldsymbol{\theta}^{(6)}}(\boldsymbol{x}^{(6)}_{su})$ needs five iterations only, and the rest necessitate a significantly smaller number of iterations too. For the Burger's equation, $f^m_{\boldsymbol{\theta}^{(1)}}(\boldsymbol{x}^{(1)}_{su})$

undergoes 140 iterations, while each of $f^m_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{x}^{(i)}_{su})$ for i = 2, 4, 5, 8, and 9 requires only one. An automatic evaluation of $f^m_{\boldsymbol{\theta}^*}(\boldsymbol{x}_{te})$ on $\boldsymbol{D}_{te}$ follows the meta-training for computing the RMSE of the MPINNs.

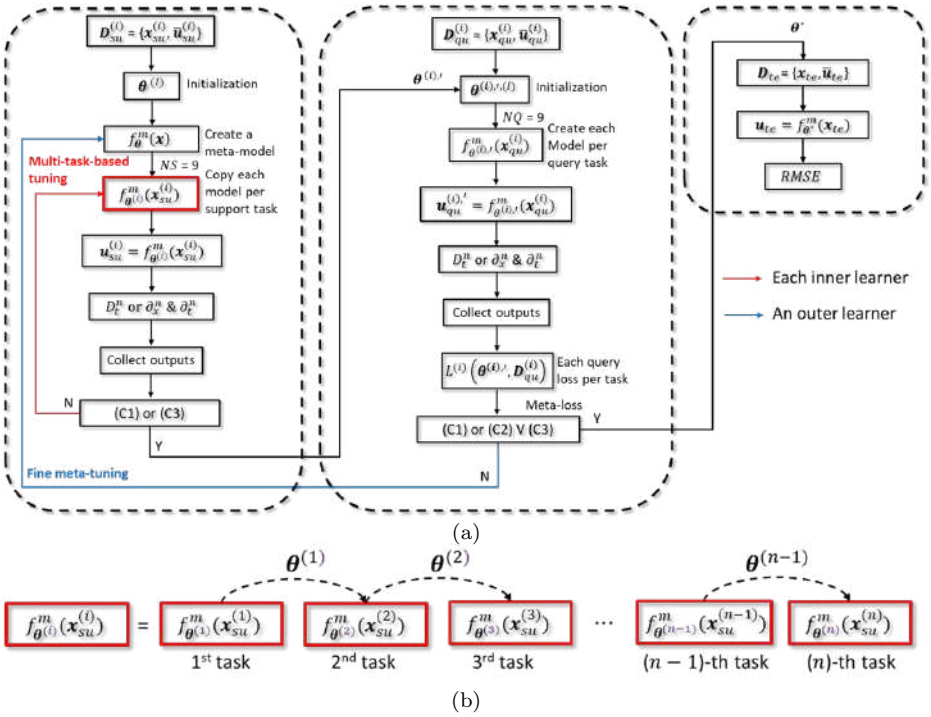


Figure 6. Schematic MPINN algorithm with the dual learners based on the MAML. (a) The meta-tuning and meta-test phases for the MPINN. (b) The application of the transfer learning to $f^m_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{x}^{(i)}_{su})$.

4.2. Experiments for the comparative performance evaluation of the PINNs and the MPINNs

In this section, the experimental results of the IVPs by the PINNs and the MPINNs are discussed. Figure 7 and Table 5 show that the MPINN decreases the overall RMSE and the local RMSE in (III) of the Duffing oscillator problem, respectively, by 98.84 % and 98.83 % relative to the PINN. The former reduces the overall RMSE and the local RMSE at t = 45, respectively, by 79.81% and 85.89 % for the Burger's equation, as shown in Figure 8 and Table 6, relative to the latter. The meta-training samples both examples in the same manner shown in (a) and (c) of Figure 2.

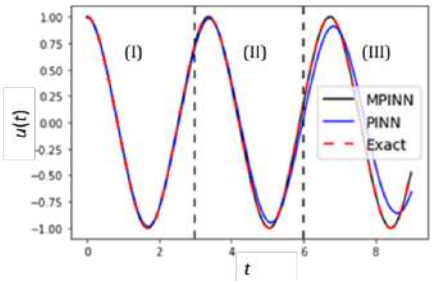


Figure 7. Comparison of the convergence of the solutions to the IVP of the Duffing oscillator.

Table 5. Comparison of the RMSEs of the IVP of the Duffing oscillator

Solver	RMSE			
	(I)	(II)	(III)	Overall
PINN	1.673×10^{-2}	5.967×10^{-2}	1.697×10^{-1}	1.043×10^{-1}
MPINN	1.854×10^{-4}	7.842×10^{-4}	1.979×10^{-3}	1.212×10^{-3}

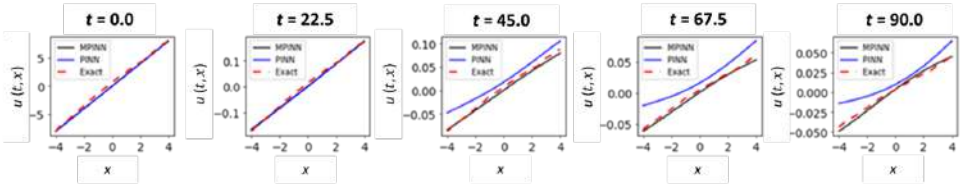


Figure 8. Comparison of the convergence of the solutions to the IVP of the Burger's equation.

Table 6. Comparison of the RMSEs of the IVP of the Burger's equation

Solver	RMSE					
	$t = 0$	$t = 22.5$	$t = 45$	$t = 67.5$	$t = 90$	Overall
PINN	1.057×10^{-3}	2.198×10^{-3}	2.277×10^{-2}	2.247×10^{-2}	1.616×10^{-2}	1.293×10^{-2}
MPINN	5.937×10^{-4}	3.651×10^{-3}	3.213×10^{-3}	1.993×10^{-3}	3.605×10^{-3}	2.611×10^{-3}

Based on the experimental data, we conclude that the MPINNs with reinitialization capability effectively solves the IVPs presented in the examples, whereas the PINNs with no fine adjustment function struggle to solve them.

5. Adaptability and correlation analyses

To showcase the adaptability of initial parameters, we first examine the evolution of the initial values of $\boldsymbol{\theta}^{[l]}$ of $f_{\boldsymbol{\theta}}^{m}(\boldsymbol{x})$ before (either the VS or the normal Xavier is implemented for pre-meta-training) and after the fine reinitialization by the meta-training and then sort out hidden layers that play a key role in the reinitialization

and adaptation (see Figure 9 and 10). From the data, there is no doubt that $\boldsymbol{\theta}^{(0)}$ initiates it for both IVP problems.

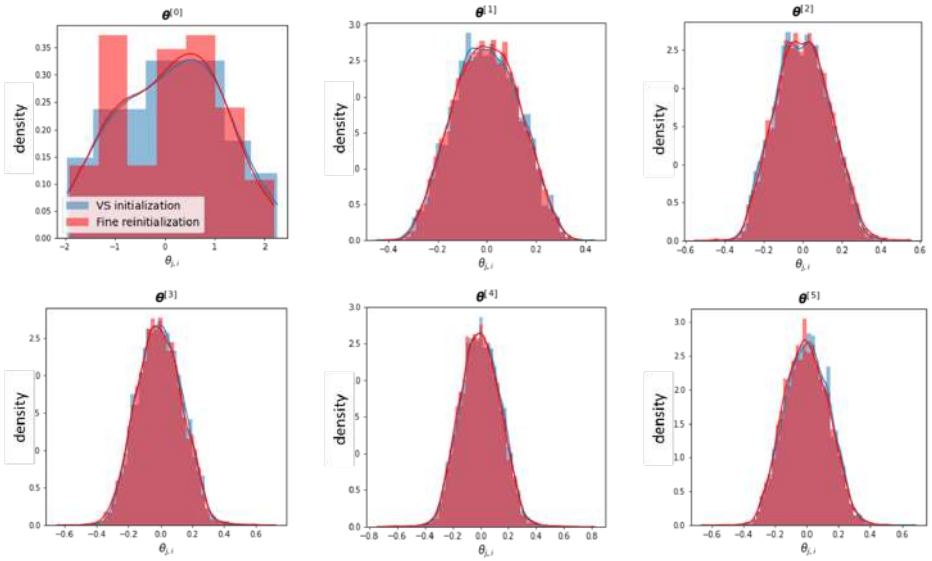


Figure 9. Evolution of the distribution of the initial values of $\boldsymbol{\theta}^{[l]}$ of the Duffing oscillator.

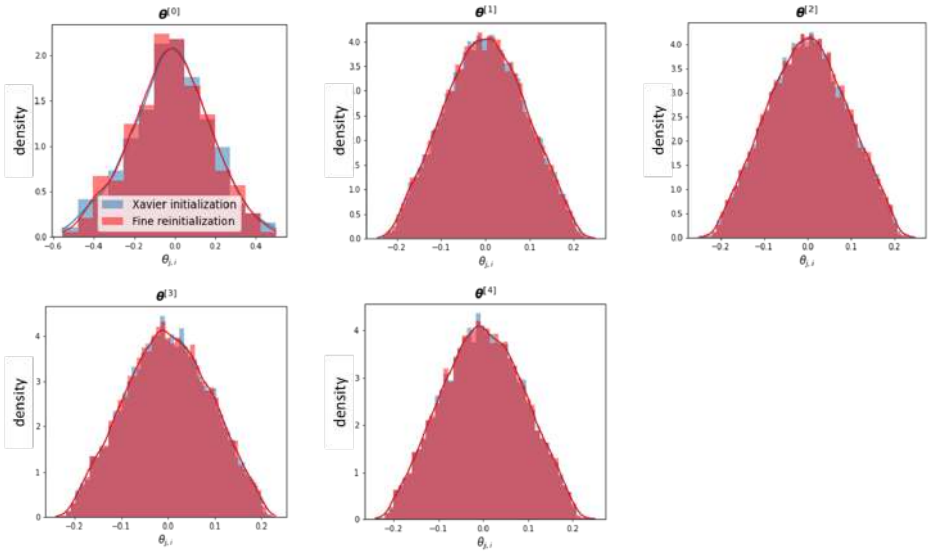


Figure 10. Evolution of the distribution of the initial values of $\boldsymbol{\theta}^{[l]}$ of the Burger's equation.

This section investigates how the initial values of $\boldsymbol{\theta}^{[0]}$s are related to the outputs of the examples. Additionally, as a supplementary to Figure 5, changes in the initial values of $\boldsymbol{\theta}^{[0]}$s before and after the application of the meta-learning are highlighted. As for the issues, correlation analyses are conducted using Kendall's

rank-order correlation coefficient. To make heatmaps for the analyses more readable, in addition to different colors, a square factor with different sizes is added to them. The size of a square in a correlation matrix generally corresponds to the strength of the correlation, with a larger size indicating a stronger correlation. The color of the square indicates the sign of the correlation, with blue and red typically representing positive and negative correlations, respectively. The depth of the color, either darker blue or darker red, represents the strength of the correlation with the square size.

Figure 11 displays the sampling of input coordinates for $u(t,x)$ that is used for correlation analysis between $\theta_{j,i}^{[0]}$ and $u(t,x)$. 256 samples are collected at random. In Figures 12-15, an x-axis or y-axis is graduated from left to right or top to bottom, respectively, and the graduations are labeled with $\theta_{j,i}^{[0]}$ values where necessary. To label them, the Duffing oscillator begins with $i=1$, then selects $j=1,2,\ldots,64$. The Burger's equation begins with $i=1$, then selects $j=1,2,\ldots,128$, and this process repeats for $i=2$. For the first example, Figure 12.(a) and 12.(b) show the heatmaps of the correlation between $\boldsymbol{\theta}^{[0]}$ initialized, respectively, by the VS (before the use of the meta-learning) and the re-initializer (it means the whole meta-learning process) and the output, $u(t)$. The comparative analysis of both heatmaps demonstrates that they become less correlated during the meta-learning. The PINN with the two-point initial condition is sensitive to the initial value of $\boldsymbol{\theta}^{[0]}$, and the mitigation of the sensitivity, as inferred from the lower correlation between the initial value of $\boldsymbol{\theta}^{[0]}$ and $u(t)$, improves the model accuracy. This suggests that the meta-learning effectively finds the optimal initial parameters, which leads to the improved solution accuracy in the presence of the less sensitivity to $u(t)$, as the elements of $\boldsymbol{\theta}^{[0]}$ are updated individually. It can be explained by comparing Figure 14.(a) with 14.(b) which depict the correlations between $\boldsymbol{\theta}^{[0]}$s before and after the fine reinitialization. In the case of the Burger's equation (see Figure 13.(a) and 13.(b)), unlike the Duffing oscillator, the initial value of $\boldsymbol{\theta}^{[0]}$ and the output, $u(t,x)$ become more correlated during the meta-learning. This increased sensitivity leads to the improved model accuracy, as the model captures more complex relationships between x and $u(t,x)$. It seems that the increase in the model accuracy due to the application of the meta-learning comes at the cost of the decreased robustness. The reason for this appears to be the integration of relatively intricate geometric domain information into the initial condition with a large number of samples used for discretization, making it difficult to decrease the sensitivity. However, it is imperative to be cautious, as this could lead to a less robust model that is vulnerable to the risk of overfitting and noise or inaccuracies in new input data, before finding an acceptable trade-off between robustness and model accuracy.

Figure 14 and 15 show the heatmaps of the correlation between $\boldsymbol{\theta}^{[0]}$s initialized, respectively, by the VS and the Xavier initializers (before the use of the meta-learning) and the re-initializer for both examples. The Duffing oscillator (see Figure 14.(a) and 14.(b)) demonstrates that the correlation becomes less during the meta-learning. From the data, it is clear that the re-initializer updates the elements of $\boldsymbol{\theta}^{[0]}$ independently in different directions and with different increments compared to the VS used in the pre-meta-learning. The observation supports Figure 5 well. For the case of the Burger's equation, in contrast to the former, $\boldsymbol{\theta}^{[0]}$s become more correlated as shown in Figure 15.(a) and 15.(b) which suggests that the elements of $\boldsymbol{\theta}^{[0]}$ are updated in an interdependent manner in the application of the re-initializer.

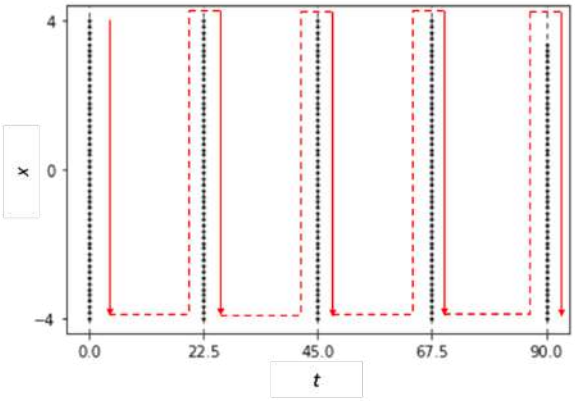


Figure 11. 256 samples for labeling the y-axis of Figure 13.(a) and 13.(b) with (t, x) top to bottom: first select $t = 0$, then select x between 4 and -4, and repeat this for t on the dashed lines (see the direction of arrows).

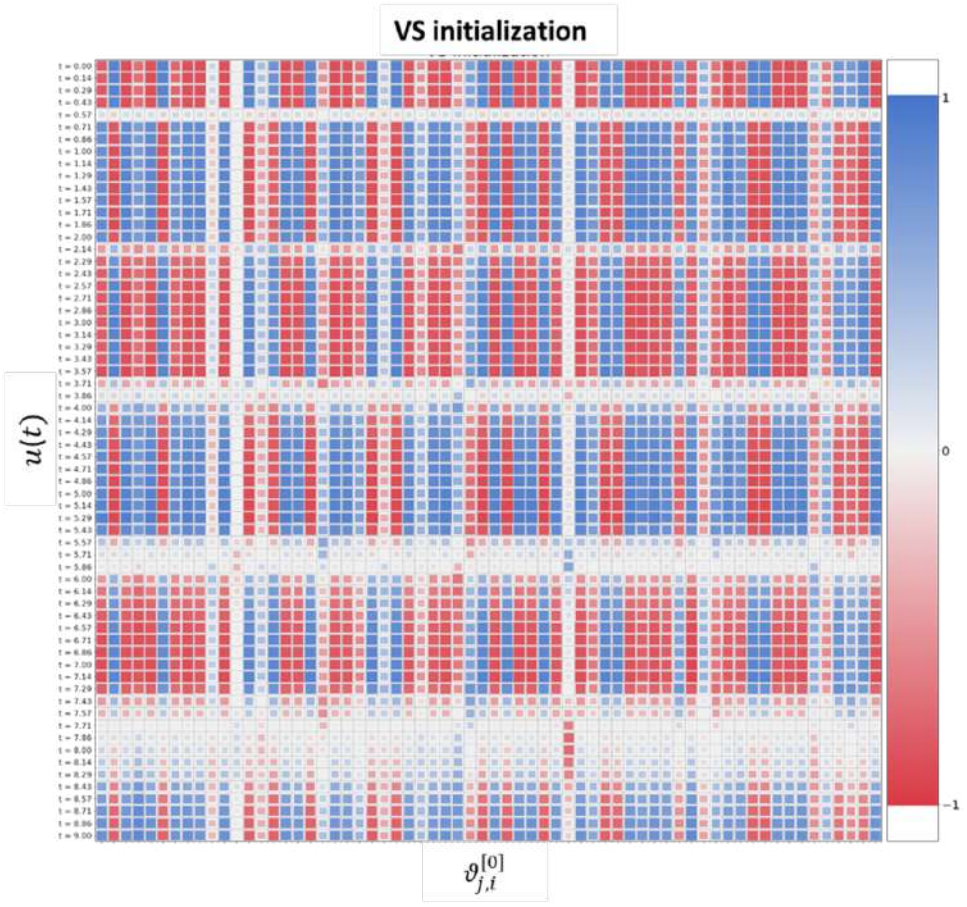


Figure 12. (a) Correlation analysis between $\boldsymbol{\theta}^{[0]}$ initialized in the last meta-iteration and $u(t)$ for the Duffing oscillator.

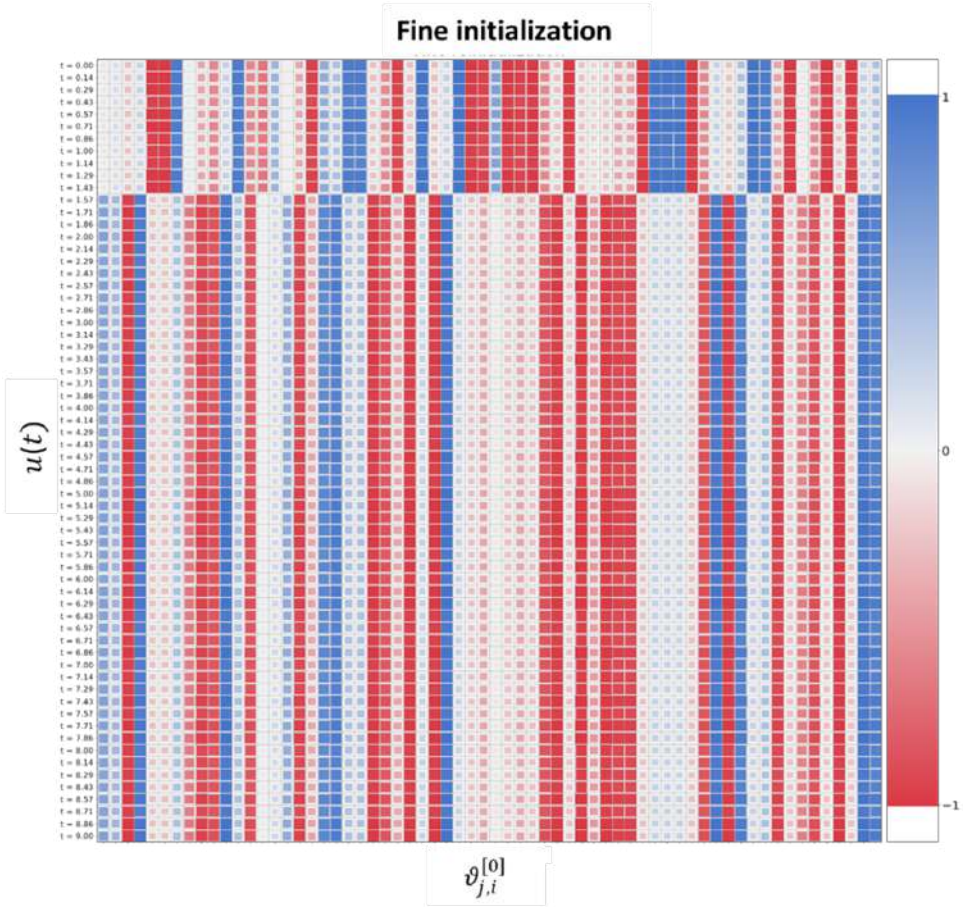


Figure 12. (b) Correlation analysis between $\boldsymbol{\theta}^{[0]}$ initialized in the last meta-iteration and $u(t)$ for the Duffing oscillator.

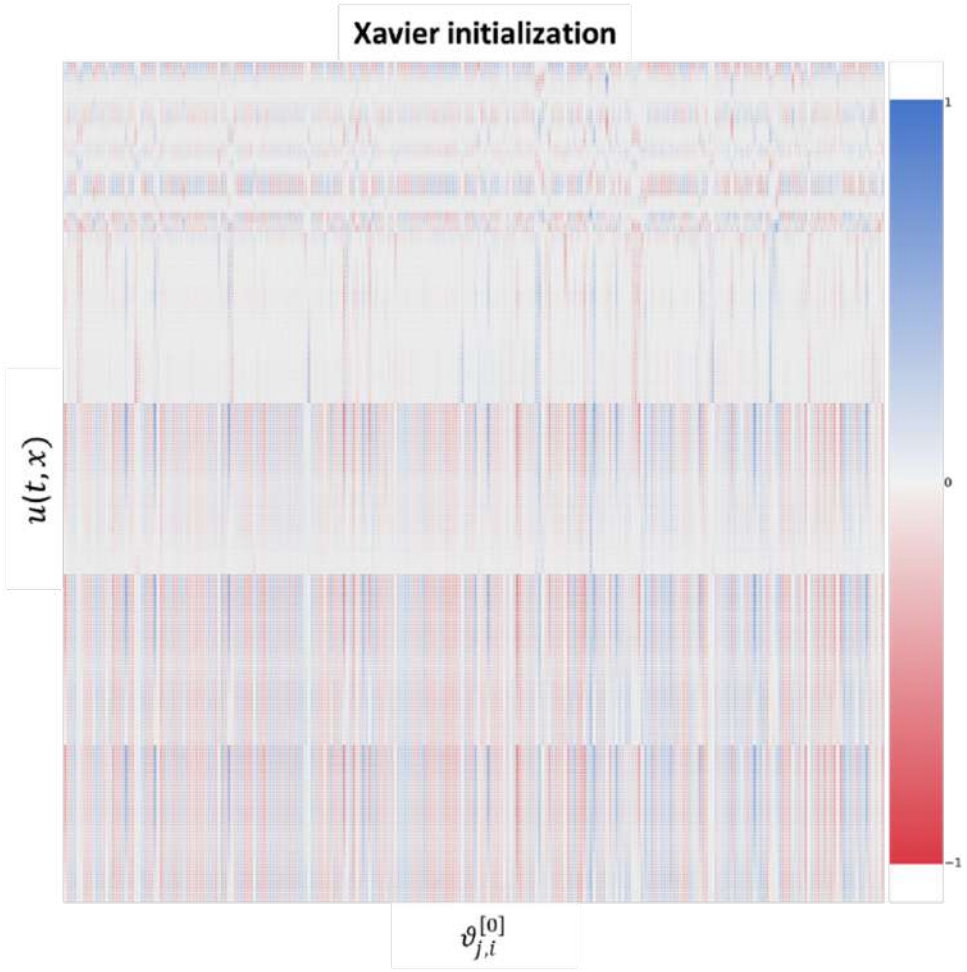


Figure 13. (a) Correlation analysis between $\boldsymbol{\theta}^{[0]}$ initialized by the Xavier and $u(t,x)$ for the Burger's equation.

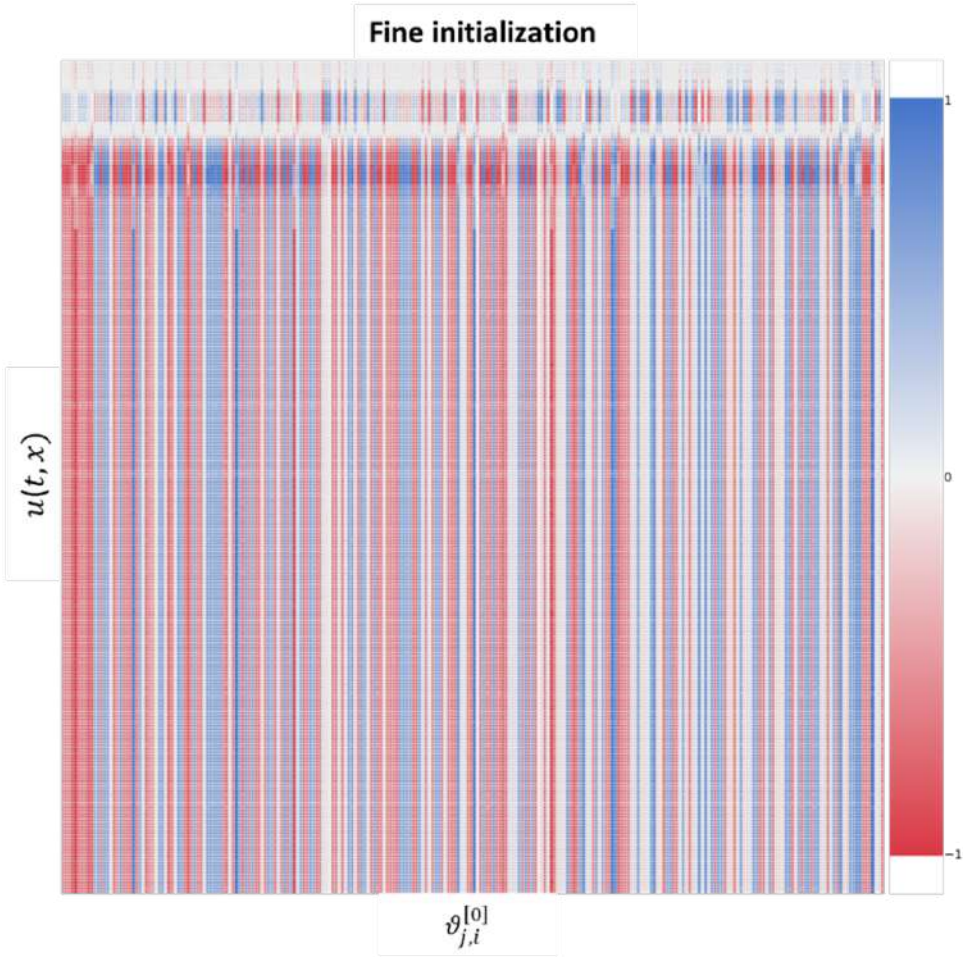


Figure 13. (b) Correlation analysis between $\boldsymbol{\theta}^{[0]}$ initialized in the last meta-iteration and $u(t, x)$ for the Burger's equation.

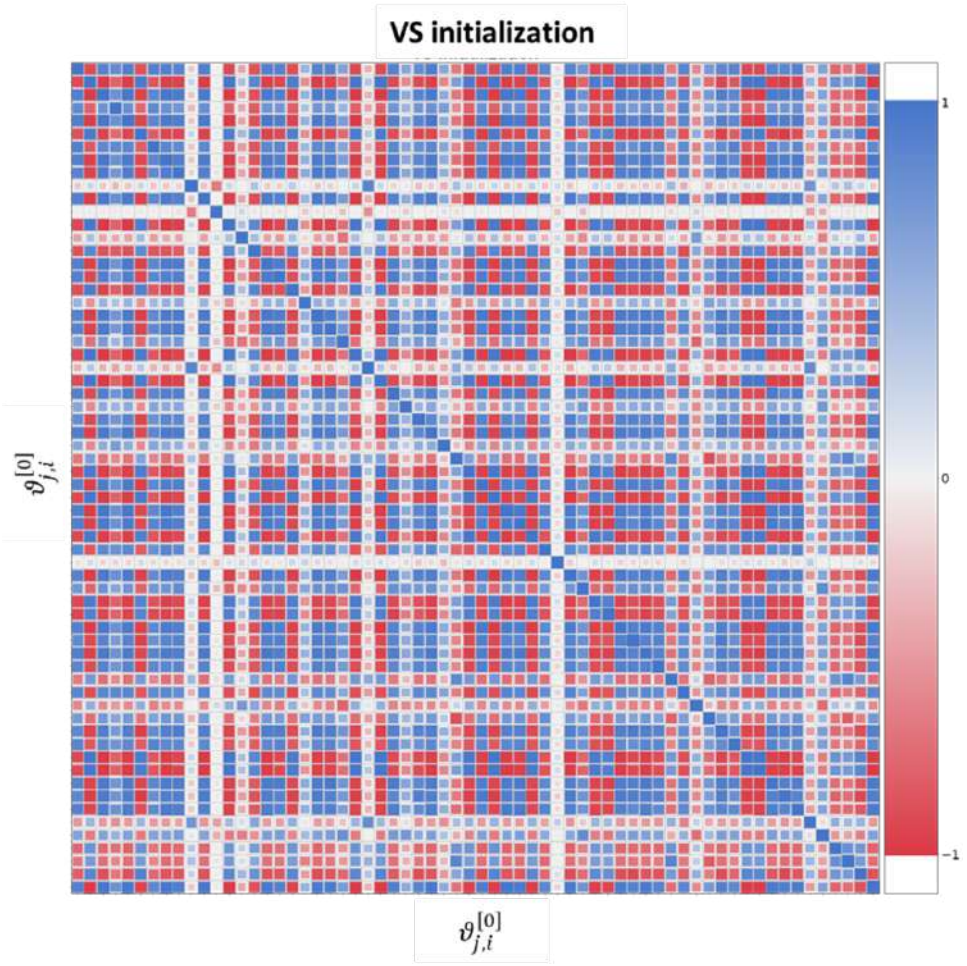


Figure 14. (a) Correlation analysis between $\boldsymbol{\theta}^{[0]}$s initialized by the VS for the Duffing oscillator.

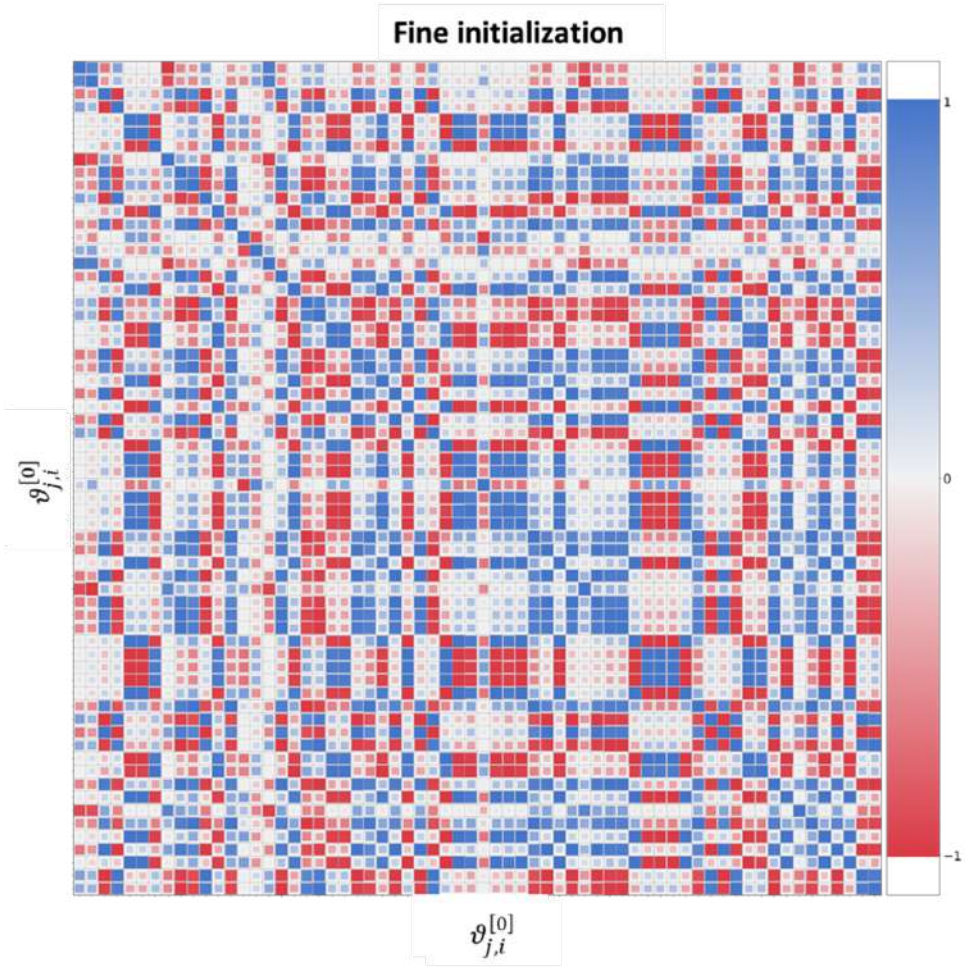


Figure 14. (b) Correlation analysis between $\boldsymbol{\theta}^{[0]}$s initialized in the last meta-iteration for the Duffing oscillator.

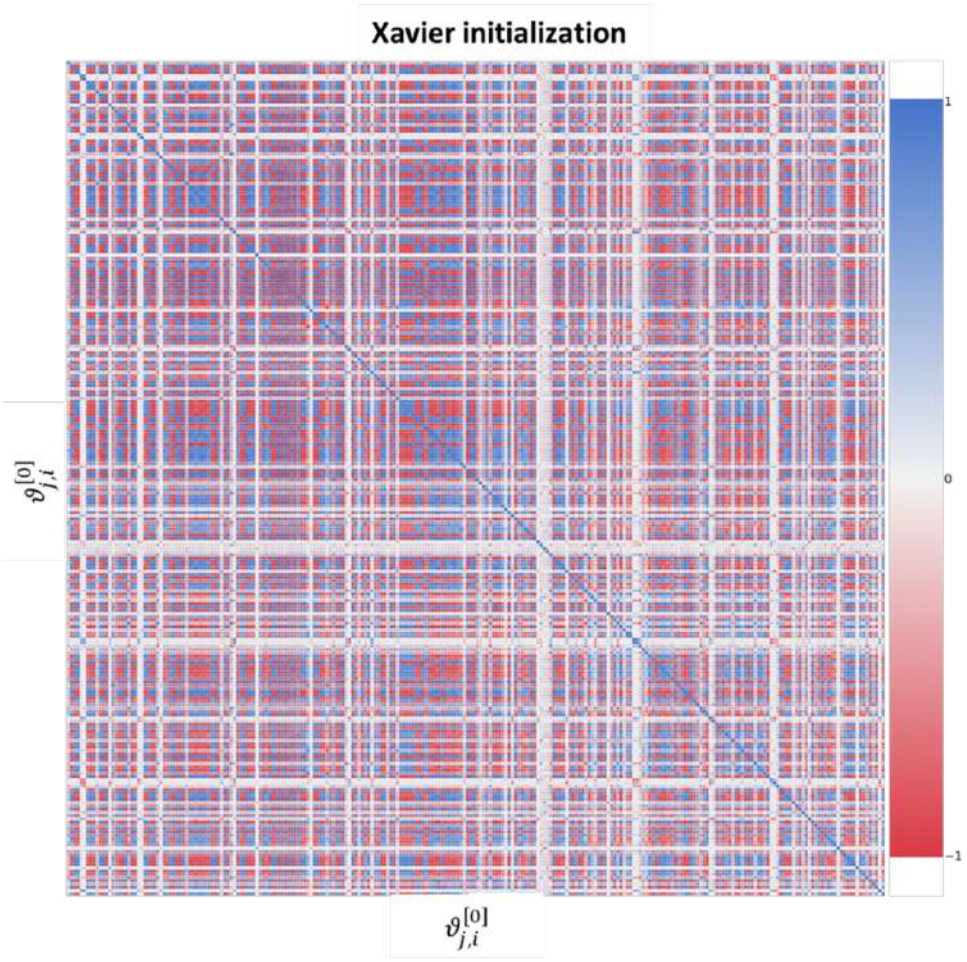


Figure 15. (a) Correlation analysis between $\boldsymbol{\theta}^{[0]}$s initialized by the Xavier for the Burger's equation.

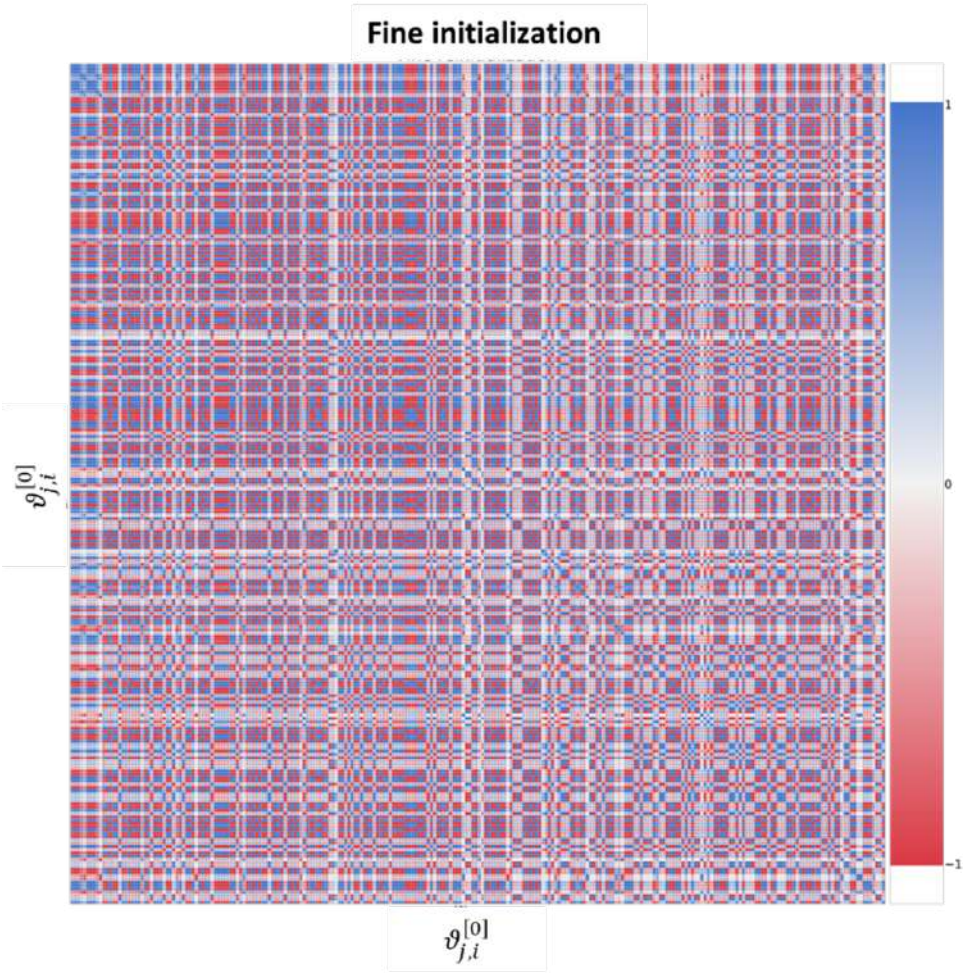


Figure 15. (b) Correlation analysis between $\boldsymbol{\theta}^{[0]}$s initialized in the last meta-iteration for the Burger's equation.

6. Conclusion

IVPs, which have significant importance in science and engineering and are associated with various real-world challenges, need for their accurate numerical solution as an essential task. In this study, we deal with the examples that are very simple but reflect the characteristics of IVPs well. The case studies are devised to encompass both scenarios: one where the output repeats itself or other similar patterns (the Duffing oscillator) and another where there is substantial variation in the gradient of the output with t (the Burger's equation). Through the scenarios, in addition to the precise numerical solutions of the IVPs, it is possible to observe a distinct dissimilarity in the relation between the solutions with high accuracy and the model's sensitivity to them when the fine reinitialization is implemented.

The PINNs with a single learner struggle to find accurate solutions to the IVPs. As evidence of this, the experimental data demonstrate the large RMSEs. It implies

that the use of PINNs in science and engineering is limited because, unlike BVP and IBVP, IVPs may necessitate fine reinitialization. In response to the limitation, we propose the MPINN with the dual learners based on the MAML. The collaboration of the multiple number of inner learners and the outer learner overcomes the limitation of the PINNs with no fine re-initializer, as it enables the optimal reinitialization of $\boldsymbol{\theta}$ and gives the solutions with higher precision. In this process, $\boldsymbol{\theta}^{[0]}$ initiates the fine reinitialization for both IVPs. The findings show how closely the solution convergence of the IVPs is coupled to the initial values of $\boldsymbol{\theta}^{[0]}$s.

For the Duffing oscillator, the meta-learning decreases the sensitivity of the initial value of $\boldsymbol{\theta}^{[0]}$ to $u(t)$ by updating its components in the different manner, leading to the highly accurate solution. On the contrary, in the case of the Burger's equation, updating the components of $\boldsymbol{\theta}^{[0]}$ in the interdependent manner increases the sensitivity of the meta-model to $u(t, x)$. The findings from the correlation analyses hint that the application of the meta-learning process alone for solving the PDE may be insufficient to optimize both issues at the same time, despite the improvement in the model accuracy. Applying joint optimization to the MPINN, which may reset the correlation between the initial parameters and the output, could provide a double-purpose solution. Regarding savings in computation cost, the use of the transfer learning between $f^m_{\boldsymbol{\theta}^{(i)}}(\boldsymbol{x}^{(i)}_{su})$ contributes to the reduction of the number of iterations required for the pre-meta-training of each inner learner.

7. Future works

A lot of progress has been made in self-tuning neural networks through joint optimization involving neural network architecture search and hyperparameters. It could be imperative to apply the optimization method to the MPINN for achieving both higher model accuracy and robustness. To realize it, our upcoming study will develop a hypergradient descent-based MPINN. The optimal trade-off between higher model accuracy and lower sensitivity using an event-triggered mechanism to activate a hypergradient descent process could be achieved.

Acknowledgements

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by King Mongkut's Institute of Technology Ladkrabang (KMITL) of Thailand (grant number: 2566-02-01-007).

References

[1] R. Zhang, Y. Liu and H. Sun, *Physics-informed multi-lstm networks for metamodeling of nonlinear structures*, Computer Methods in Applied Mechanics and Engineering, 2020, 369, 113226.

[2] S. Dong and Z. Li, *Local extreme learning machines and domain decomposition for solving linear and nonlinear partial differential equations*, Computer Methods in Applied Mechanics and Engineering, 2021, 387, 114129.

[3] X. Meng, Z. Li, D. Zhang and G. E. Karniadakis, *PPINN: Parareal physics-informed neural network for time-dependent pdes*, Computer Methods in Applied Mechanics and Engineering, 2020, 370, 113250.

[4] R. Zhang, Y. Liu and H. Sun, *Physics-guided convolutional neural network (phycnn) for data-driven seismic response modeling*, 2019.

[5] M. De Florio, E. Schiassi and R. Furfaro, *Physics-informed neural networks and functional interpolation for stiff chemical kinetics*, Chaos: An Interdisciplinary Journal of Nonlinear Science, 2022, 32(6), 063107.

[6] J. Sirignano and K. Spiliopoulos, *Dgm: A deep learning algorithm for solving partial differential equations*, Journal of Computational Physics, 2018, 375, 1339–1364.

[7] M. Raissi, P. Perdikaris and G. Karniadakis, *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations*, Journal of Computational Physics, 2019, 378, 686–707.

[8] Y. Zhu, N. Zabaras, P.-S. Koutsourelakis and P. Perdikaris, *Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data*, Journal of Computational Physics, 2019, 394, 56–81.

[9] L. Sun, H. Gao, S. Pan and J.-X. Wang, *Surrogate modeling for fluid flows based on physics-constrained deep learning without simulation data*, Computer Methods in Applied Mechanics and Engineering, 2020, 361, 112732.

[10] A. D. Jagtap, E. Kharazmi and G. E. Karniadakis, *Conservative physics-informed neural networks on discrete domains for conservation laws: Applications to forward and inverse problems*, Computer Methods in Applied Mechanics and Engineering, 2020, 365, 113028.

[11] L. Sun and J.-X. Wang, *Physics-constrained bayesian neural network for fluid flow reconstruction with sparse and noisy data*, Theoretical and Applied Mechanics Letters, 2020, 10(3), 161–169.

[12] N. Geneva and N. Zabaras, *Modeling the dynamics of PDE systems with physics-constrained deep auto-regressive networks*, Journal of Computational Physics, 2020, 403, 109056.

[13] M. Baymani, A. Kerayechian and S. Effati, *Artificial neural networks approach for solving stokes problem*, Applied Mathematics, 2010, 1 No. 4, 288–292.

[14] H. Gao, L. Sun and J.-X. Wang, *Phygeonet: Physics-informed geometry-adaptive convolutional neural networks for solving parameterized steady-state pdes on irregular domain*, Journal of Computational Physics, 2021, 428, 110079.

[15] Y. Ma, X. Xu, S. Yan and Z. Ren, *A preliminary study on the resolution of electro-thermal multi-physics coupling problem using physics-informed neural network (PINN)*, Algorithms, 2022, 15(2).

[16] C. Rao, H. Sun and Y. Liu, *Physics-informed deep learning for incompressible laminar flows*, Theoretical and Applied Mechanics Letters, 2020, 10(3), 207–212.

[17] S. A. Niaki, E. Haghighat, T. Campbell et al., *Physics-informed neural network for modelling the thermochemical curing process of composite-tool systems during manufacture*, Computer Methods in Applied Mechanics and Engineering, 2021, 384, 113959.

[18] Q. Zhu, Z. Liu and J. Yan, *Machine learning for metal additive manufacturing: predicting temperature and melt pool fluid dynamics using physics-informed neural networks*, Computational Mechanics, 2021, 67, 619–635.

[19] L. Lu, X. Meng, Z. Mao and G. E. Karniadakis, *Deepxde: A deep learning library for solving differential equations*, SIAM Review, 2021, 63(1), 208–228.

[20] H. Lee and I. S. Kang, *Neural algorithm for solving differential equations*, Journal of Computational Physics, 1990, 91(1), 110–131.

[21] A.J. Meade and A. Fernandez, *The numerical solution of linear ordinary differential equations by feedforward neural networks*, Mathematical and Computer Modelling, 1994, 19(12), 1–25.

[22] I. Lagaris, A. Likas and D. Fotiadis, *Artificial neural networks for solving ordinary and partial differential equations*, IEEE Transactions on Neural Networks, 1998, 9(5), 987–1000.

[23] H. Robinson, S. Pawar, A. Rasheed and O. San, *Physics guided neural networks for modelling of non-linear dynamics*, Neural Networks, 2022, 154, 333–345.

[24] Y. Peng, P. A. Flach, C. Soares and P. Brazdil, *Improved dataset characterisation for meta-learning*, in *Discovery Science* (Edited by S. Lange, K. Satoh and C. H. Smith), Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, 141–152.

[25] Y. Bengio, S. Bengio and J. Cloutier, *Learning a synaptic learning rule*, in *IJCNN-91-Seattle International Joint Conference on Neural Networks*, ii, 1991, 969 vol.2–.

[26] S. Bengio, Y. Bengio, J. Cloutier and J. Gecsei, *On the Optimization of a Synaptic Learning Rule*, Routledge, London, 2014, 265–287.

[27] S. Hochreiter, A. S. Younger and P. R. Conwell, *Learning to learn using gradient descent*, in *Artificial Neural Networks – ICANN 2001* (Edited by G. Dorffner, H. Bischof and K. Hornik), Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, 87–94.

[28] A. Younger, S. Hochreiter and P. Conwell, *Meta-learning with backpropagation*, in *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, 3, 2001, 2001–2006 vol.3.

[29] C. Finn, P. Abbeel and S. Levine, *Model-agnostic meta-learning for fast adaptation of deep networks*, in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, JMLR.org, 2017, 1126–1135.

[30] X. Liu, X. Zhang, W. Peng et al., *A novel meta-learning initialization method for physics-informed neural networks*, Neural Computing and Applications, 2022, 34, 14511–14534.

[31] F. d. A. Belbute Peres, Y.-f. Chen and F. Sha, *HyperPINN: Learning parameterized differential equations with physics-informed hypernetworks*, in *The Symbiosis of Deep Learning and Differential Equations*, 2021.

[32] A. F. Psaros, K. Kawaguchi and G. E. Karniadakis, *Meta-learning pinn loss functions*, Journal of Computational Physics, 2022, 458, 111121.

[33] M. Penwarden, S. Zhe, A. Narayan and R. M. Kirby, *A metalearning approach for physics-informed neural networks (PINNs): Application to parameterized pdes*, Journal of Computational Physics, 2023, 477, 111912.

[34] M. Huisman, J. N. van Rijn and A. Plaat, *A survey of deep meta-learning*, Artificial Intelligence Review, 2021, 54, 4483–4541.

Cluster Computing
https://doi.org/10.1007/s10586-024-04559-4

Modified clustering and incentivized stable CH selection for reliable VANET communication

Ashish Kumari[1,2] · Shailender Kumar[1] · Ram Shringar Raw[3]

Received: 31 January 2024 / Revised: 18 March 2024 / Accepted: 7 May 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
Modern smart cities rely on Vehicular Ad-hoc Networks (VANETs) for real-time traffic safety warnings and collision avoidance. However, high mobility, variable driving patterns, and urban contexts in VANET require optimal clustering and stable Cluster Head (CH) selection for reliable communication. Problem statement: Due to dynamic mobility in VANET, maintaining the CH stability and reliable communication is still a challenge. In high-density urban environments, the risk of communication disruptions increases due to the presence of a large number of vehicles competing for limited communication resources. Traditional graph-based suboptimal clusters lead to suboptimal communication reliability and disruptions compared to hypergraph based VANET. Such a modelling leads to reduced overhead and increased packet delay. Proposed Methodology: To overcome these issues of varying mobility in VANET and unstable CH selection, a new CH selection framework with an Adaptive Neighboring Behavior-Incentivization (ANB-I) process for stable and reliable CH selection is introduced. The selection of the CH based on Multi-criteria Multi-decision (MCMD) making of three metrics: Adaptive Neighborhood Degree, Adaptive Link Lifetime (Adaptive -LLT), and Adaptive Relative Average Speed. Results: The suggested approach prioritizes CHs with normal behaviours and longer link lifetimes to balance stability and reliability. The proposed model is tested on MATLAB software and SUMO simulations of 100 (sparse) and 1000 (dense) vehicle densities. The proposed model achieves a 75% CH stability with three optimal numbers of clusters, outperforming previous works by 4.17% improvement. Conclusion: A key aspect of the proposed CH selection is the incorporation of dynamic behaviour of driver's estimation based on the concept of incentive process. By prioritizing CHs with longer link lifetimes and stable driving patterns, the proposed model seeks to mitigate disruptions and enhance the throughput (overhead) of the VANET network.

Keywords VANET · Adaptive Neighboring Behavior-Incentivization CH selection · SUMO · MATLAB and CH stability

1 Introduction

1.1 Overview

Communication between the vehicles is done with the help of a VANET, a modified form of a Mobile Ad-hoc Network (MANET). It is a significant part of the intelligent transportation system, in which vehicles have different communication modes such as vehicle to structure, vehicle to vehicle, and anything that comes in contact [1]. The standards for the communication that are used are set by the IEEE 802.11p for short-range communication and wireless access in vehicular environments. The functionality of VANET relies on network nodes, which experience frequent changes in position due to the high mobility of vehicles. This results in significant network instability and scalability issues, as depicted in Fig. 1. The node latency in VANET is greater than in MANET, leading to a loss in communication quality and causing delays in timely packet delivery [1].

✉ Shailender Kumar
shailenderkumar@dce.ac.in

[1] Delhi Technological University, New Delhi 110042, India

[2] Maharaja Surajmal Institute of Technology, C-4, Janakpuri, New Delhi 110058, India

[3] NSUT, East Campus (Formerly AIACTR), New Delhi 110031, India

Published online: 07 June 2024

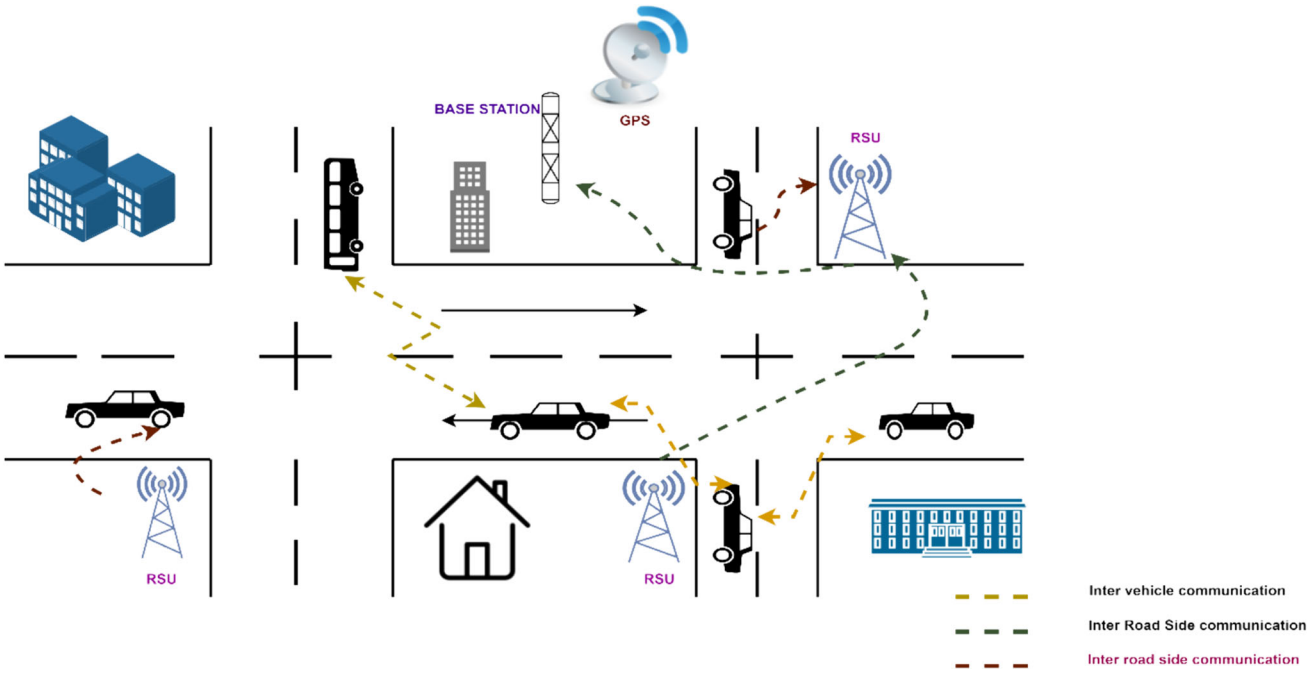


Fig. 1 Basic diagram showing VANET architecture

Within the domain of VANETs, the network exhibits a highly dynamic nature, characterized by the unpredictable and unusual movement of nodes (vehicles), resulting in frequent alterations to the network structure. The constant change and variability present in this situation create difficulties in maintaining stable communication and achieving efficient transmission [3, 4]. More precisely, the frequent rebuilding of the network leads to unreliable connections among vehicles and neighboring vehicles, thereby reducing the overall effectiveness of transmitting information in VANETs [3]. Hence, it is crucial to address the issues of preserving connection quality and reducing the frequency of group reorganization to advance the development of VANETs. To address these issues, researchers have developed algorithms based on the concept of clustering [5]. Most of the clustering methods efficiently classify vehicles into several clusters by utilizing characteristics such as inter-vehicle distance and relative speed. Each cluster consists of a CH and Cluster Members (CMs) that are deliberately chosen. Clustering of vehicles in VANET enhances the development of controllable and expandable VANET clusters, promoting a more equitable allocation of network burden.

The formation of a dynamic virtual backbone by CH plays a pivotal role in handling essential functions like routing and channel allocation. Particularly, in the case of transferring an ongoing vehicle data session from one Roadside Unit (RSU) to another, known as handover, challenges arise, especially for fast-moving vehicle groups. The conventional approach of individual vehicle handovers results in significant signalling overhead between vehicles and RSUs, leading to suboptimal performance. Conventional clustering techniques, such as K-means, have difficulties in effectively managing the changing characteristics of evolving VANETs. Whereas, another method of clustering, spectral clustering prioritizes data connectedness above geometric proximity, offering a semi-convex method that is better suited for the changing graph dynamics of VANETs [10, 13–15]. However, in this study, a hypergraph-based mTTM (Modified Tensor Trace Matrix) model is implemented for clustering. Whereas, the stability of a cluster is contingent on the frequency change in CM and in turn frequency of CH reselection, both of which are essential parameters in CH selection mechanisms. Connecting the stability to the space headway and connectivity by [20], the vehicle stability within a lane is intricately tied to its degree of connectivity in that lane. Higher connectivity implies greater stability, a concept intuitively linked to vehicle density and space headway [20]. Considering this concept, unlike traditional CH metrics, the concepts of dynamic behaviour and relativity are introduced in this study for a novel CH selection metric.

In the presented study, the concept of relativity plays a crucial role in the CH selection mechanism, particularly through the innovative Relative Neighboring Behavior-Incentivization process. In this study, relativity is introduced for the adjustment of CH metrics based on the dynamic behaviour and context of vehicles within the vehicular clusters inspired by [4, 7, 9]. The need for relativity arises from the diverse and dynamic nature of driving

patterns exhibited by vehicles in different scenarios [20], which can significantly impact the stability and reliability of the chosen CH. Relativity, therefore, serves as a dynamic mechanism to align the CH selection parameters with the real-time behaviour of vehicles. It allows the incentivization process to adapt to the varying conditions and driving styles observed in the vehicular network.

1.2 Challenges

There is a recent advancement in utilizing spectral clustering for VANET clustering [13–15] CoVEG (cluster-based VANET-oriented evolving graph) [23, 27]. For CH selection in VANET, most of the studies highlight the importance of considering factors such as trust [3, 11, 15, 22], relative speed [15, 22, 23, 27] and such. However, there is a need for research on prioritising the selection of CH. The identified problems in the recent research are,

- With high mobility, in areas like Connaught in Delhi [22], it becomes challenging to prioritize the selection of CH which shows usual behaviour or aggression as explained in [4, 7, 9] and maintains the longer link lifetimes. Hence, there is a need to consider adaptive and responsive CH selection mechanisms.
- Most of the studies concentrate on VANET analysis in sparse contexts like roadways. However, the transportation network functions differently in an urban setting due to the high vehicle density. The network becomes increasingly challenging as more connections and disconnections occur simultaneously. A super network is what is used when more than two cars are connected at once. The dependence of a vehicle on all of its Neighbors is not considered by traditional network mapping as a two-node graph [22].
- In traditional spectral clustering, the adjacency matrix is a key component used to capture the relationships or similarities between vehicles. However, traditional spectral clustering may result in less detailed community structures within the generated clusters. There is a need for transformation that allows for the determination of the dependency of the vehicle's feature matrix on social connectivity [29].

1.3 Contributions of the paper

Motivated by these challenges, the proposed method for CH selection metrics is introduced with the concept of relativity and incentivized dynamic behaviour to select a stable CH. The main contributions of the paper are,

- The paper introduces an innovative approach to representing VANETs as hypergraphs. It leverages Eigen Heuristics as a means to determine the optimal number of clusters before initiating the clustering process.
- A dynamic driving behaviour assessment mechanism for CH selection based on incentives is proposed. It considers the dynamic driving patterns of vehicles within clusters, distinguishing between usual and unusual patterns of velocity.
- The paper employs three key metrics, namely Adaptive Neighborhood Degree, Adaptive Link Lifetime, and Adaptive Relative Average Speed, to initiate the CH selection process. By incorporating an adaptive thresholding approach (Eq. 3), termed Relative Neighboring Behavior-Incentivization (ANB-I), the metrics are dynamically adjusted based on the relative behaviour of vehicles, fostering a more responsive and context-aware CH selection.
- The proposed model is simulated for CH stability over varying optimal clusters and network parameters for varying vehicle densities (100 and 1000). The metrics proposed with ANB-I is evaluated with the simple metrics over varying vehicles.

The following is the outline of this paper: The conceptualization of the problem is discussed in Sect. 3, which outlines a hypergraph approach to system modelling by making use of the Modularity Matrix and stable and secure CH selection. In Sect. 4, an overview of the proposed methodology is presented, which includes modifications to Spectral Clustering as well as parameters for selecting secure and stable CH. Finally, results and conclusion are detailed in Sect. 5 and 6 respectively.

2 Literature

2.1 Overview

The selection of stable and secure CHs in VANETs can be significantly influenced by the actions of the drivers themselves. It is difficult to select appropriate CHs due to noise and unpredictability introduced into the network connectivity by the dynamics of drivers, such as their speed, direction, and communication patterns, as will be explored in more detail below. In the following section, the challenges faced by current approaches and shedding light on the areas are discussed. The Section is divided into two parts, first discussing on the impact of driving behaviour on CH selection and other is the related works (spectral and non-eigen based clustering).

2.2 Dynamic behavior of drivers and its impact on CH selection

The dynamic behaviour of drivers plays a pivotal role in the selection of CH within vehicular networks, influencing the overall effectiveness of existing methods [20]. In this section, various approaches and their impact on CH selection are discussed. This emphasises the need for dynamic solutions that address the complexities associated with driver behaviour of a vehicle in the cluster to be elected as a CH.

The approach presented in [4] is distributed and thus more effective than existing methods for detecting aggressive driving. Intentionally aggressive driving behaviour can be uncovered through third-party observations. Although RSSI (Received Signal Strength Indicator) measurements may not be able to pick up on every instance of aggressive driving, they do give a framework for the rapid identification and alerting of nearby motorists. This gives other drivers time to respond and avoid dangerous situations. This work introduces a distributed approach for detecting aggressive driving behaviors, which can impact CH selection by influencing the dynamic behavior of vehicles within clusters. Johnson and Trivedi [5], proposed a novel system that utilizes Dynamic Time Warping (DTW) and smartphone-based sensor-fusion (gyroscope, accelerometer, video, GPS, magnetometer) to identify, recognize, and record such actions without external processing to raise awareness and promote driver safety. By combining inter-axial data from numerous sensors into a single classifier, the system offers a novel take on the field of driving pattern identification. Further aiding classification is an Euler representation of device attitude (also derived from fused data). Everything is processed entirely within the mobile device [5]. Hence, understanding driving styles and patterns can aid in selecting CHs that cater to the preferences and behavior of drivers within clusters.

Paper [7] introduced a new weighted evidence-combining strategy to fuse the contradictory information by analyzing the driving behaviour recognition parameter data. The proposed combination technique not only keeps Dempster's combination rule's great mathematical features but also takes into account all the interdependencies and effects of the evidence's body. The level of disagreement between the evidence is first quantified using a novel dissimilarity metric. In addition, uncertainty measurement and internal conflict are employed to assign relative importance to the various pieces of evidence. Reasonable aggregate results are obtained using a weighted averaging combination approach and the experimental results show that the suggested method outperforms the state-of-the-art approaches, making it a viable option for evaluating drivers' actions. Hence, by considering multiple sources of evidence, the method enhances the reliability of CH selection based on driver behavior. To strengthen the backbone of cooperative driving, Hu and Lee proposed a better vehicle clustering in [10]. The work's primary contribution is an innovative method for increasing vehicle reliability by maximizing the clustering of vehicle lifetime problems based on their related node representations, achieved with the help of the GNN model. As a result, it concludes that a GNN-based grouping algorithm that makes use of both features and graph structure outperforms the conventional approaches [10]. Kaur et al. propose that, despite the inherent instability of VANETs, these networks can provide crucial safety features during critical situations. Future guidelines for such important criteria as selecting CMs, CHs, and cluster management are the primary focus of the research. As it is possible to implant all of the managers for each intersection despite the lack of in structure and therefore cover the entire area, they concluded that the plethora of protocols used to connect VANETs just adds to the difficulty of doing so [14].

In order to foresee and prevent car accidents, Kalaivani advocated using vehicular communication technology to provide urgent alerts. This study's key contribution is that advancements in ITS technology have allowed for enhanced network safety, convenience, and longevity. As a result, the paper concludes that, in a smart transportation system, mobile communication technology is utilized to transmit urgent signals to alert vehicles in order to anticipate and prevent car accidents [12]. Hence, CH selection based on network safety considerations can benefit from advancements in ITS technology.

The paper by Hu et al. presents a method for determining the network architectures using Dirichlet processing of a Gaussian mixture model with spectral segmentation. As the study's central focus, a method for identifying community structures and applying it to the examination of many networks is presented. A novel method based on spectral clustering and DPGMM for studying community networks with no overlap is investigated. So, it can be concluded from many studies that this method works effectively in a wide range of settings. Modularity and other assessment indices are extremely useful for both synthetic networks built by computers and other types of networks in the real world. The suggested approach was evaluated against state-of-the-art techniques [16] on both real-world and simulated networks. By exploring community detection techniques, this work provides insights into network structure, which can aid in optimizing CH selection based on community structures. The author, Arkian et al., described a new three-step methodology for predicting traffic on a road segment based on the actual number of wirelessly connected vehicles. As a result, it can be said that the hybrid approach is a scalable

system that makes the most of the number of enabled vehicles travelling along a given road segment to provide an accurate estimate of the traffic flow. As a result, a novel clustering method, a clustered chain methodology, and a traffic-control approach are proposed in [17]. In the work by Miyajima et al., they suggested a model that mimics the accelerator pedal's behaviour under normal driving conditions. The key contribution of this work is a model that can learn the behaviour of the driver's pedal pattern and use it to distinguish between their driving styles. This paper concludes that driver simulations were evaluated in driver recognition trials [18] by using driving signals collected from a real car as well as a driving simulation. This study focuses on modeling and recognizing driver behaviour, which is crucial for understanding the dynamic behavior of vehicles within clusters. Driver behavior can affect vehicle dynamics and communication patterns within VANETs, influencing CH selection strategies. By modelling and recognizing driver behaviour, this work provides valuable insights into the interactions between drivers and vehicles, which can inform CH selection based on driver behavior characteristics.

Sharma et al., proposed an incentive-based reputation scheme, where nodes are first grouped according to their social features. The weighted social tie is determined to represent the strength of their social connections, and finally, the weighted social tie is updated based on any rewards or punishments that have been assigned to the nodes. If there is a tradeoff between residual energy and packet delay, or if packet latency is penalized, then the incentive is offered. To do this, IRU-mDANMF (Deep autoencoder-based non-negative matrix factorization) is created which ism DANMF, a new modularized that uses social features to determine a node's reputation (incentivized reputation update by modularized DANMF). Experiments in a variety of settings have shown that the approach outperforms the best existing options by a wide margin [19]. The incentive-based reputation scheme proposed in this work aims to enhance the reliability and trustworthiness of vehicles within the network. In VANETs, selecting reliable CHs is essential for maintaining network stability and ensuring secure communication. Inspired by this work, by incentivizing desirable behavior and penalizing undesirable behavior, this approach contributes to selecting CHs capable of effectively managing network resources and maintaining communication reliability. Zepeda et al. introduced a dynamic clustering-based driving styles classification and profile approach in their study [20], in which groups shift in reaction to external stimuli. A position-dependent dynamic clustering structure was designed in which a driver is assigned to a cluster sequence rather than a single cluster to better capture dynamic driving patterns and comprehend the driving style transition behaviours and more complex driving patterns. Zhou et al. conducted a study on the intrinsic characteristics of network traffic by examining real-world traffic trace data, and their findings show that traffic possesses the characteristics of temporal stability and geographical affinity. They model TM as multi-Gaussian distributions, which more precisely describe the actual network traffic, based on the spatial features of the traffic. To further estimate the missing traffic data, a novel matrix completion method using multi-Gaussian models was developed in [21].

Vashishth et al. presented GMMR (Gaussian Mixture Models) as a method for routing in OppIot networks in paper [25]. For the soft clustering of the network's nodes, GMMR uses Gaussian Mixture Models. The model was trained on device characteristics such as destination count, distance from the destination, device buffer occupancy, and number of successful message delivery by a device. To implement controlled message flooding, this trained model was utilized to group devices with comparable configurations. GMMR model outperforms other state-of-art models in terms of average hop count, a number of message drops, delivery probability, and overhead ratio. However, GMMR's training phase is tedious and may need to be repeated if significant changes are made to the network. In [24], Law et al. suggest a new way of learning metrics for deep supervised clustering by formulating a novel loss function. An easily calculable closed-form expression for the gradient is derived by us; its computation complexity scales linearly with the size of the training mini-batch and quadratically with the dimensionality of the representation.

From the detection of aggressive driving behaviours [4, 5] to the recognition of driving styles [18] and the prediction of traffic patterns [17, 21], these approaches demonstrate the multifaceted nature of driving behaviour based challenges faced in vehicular networks. Furthermore, innovative techniques such as community detection [16], dynamic clustering [20], and Gaussian Mixture Models for routing optimization [25] offer promising avenues for improving VANET performance and scalability. The overarching theme that emerges from this literature review is the necessity for adaptive and context-aware solutions that can effectively accommodate the dynamic nature of vehicular environments. The insights gleaned from the discussed studies lay the groundwork for further exploration and analysis of clustering methodologies within vehicular networks, underscoring the importance of dynamic and adaptive solutions in addressing the evolving challenges of vehicular communication and cooperative driving.

2.3 Related approaches for clustering in VANET

The related works discussed here are divided as spectral clustering based and non-eigen based.

2.3.1 Non-Eigen clustering based studies

The Grasshopper Optimization Algorithm (GOA) for VANETs has been used for efficient CH selection in paper [1]. However, greater node density had a detrimental effect on other state-of-art models in [1], similar to the proposed model, it addresses the challenge of optimizing cluster formation in dynamic environments. while the proposed GOA was lowering the number of clusters to achieve higher network density. To solve the problem of inefficient data delivery in urban vehicle areas, Kwon et al. [2] suggest a neighbor stability-based VANET clustering (NSVC). In addition to expediting cluster formation and reducing the frequency of CH elections, NSVC also ensures that critical information will arrive reliably in the event of an emergency. The simulation results indicate that NSVC outperforms conventional methods in terms of network performance [2]. NSVC aims to expedite cluster formation and reduce the frequency of CH elections while ensuring reliable data delivery. In paper [3], a novel clustering protocol with a new weighted formula was utilized for selecting CH; this protocol relies on trust, distance, and velocity as its three primary parameters. Clustering and selection of CH varied depending on the values of the parameters. possible that new weighted parameters could be added to an existing weighted formula for selecting CH to achieve optimal results. This protocol incorporates trust, distance, and velocity as primary parameters for CH selection, demonstrating the significance of multi-criteria decision-making in clustering algorithms. Cheng et al. [6] presented a stable clustering approach for VANETs on highways that makes use of the direction vector, the centres of vehicle density areas, and intersections to group fewer but more stable clusters. A unique CH selection algorithm and cluster maintenance algorithm were suggested, which use the relative mobility metric to mitigate the effects of vehicle distance, velocity, and maximum acceleration, thereby mitigating the effects of vehicle type and drivers' driving habits. This approach utilizes relative mobility metrics to form stable clusters, addressing challenges posed by varying vehicle distances, velocities, and driving habits. It complements our study's emphasis on stable CH selection in dynamic VANET environments.

In this paper, Iswarya and Radha proposed a secure and stable routing methodology that helps in establishing robust clustering and reduces the changing frequency between the two vehicles. The primary result of the work was the selection of a state-of-the-art CH for improving network safety and mitigating the effects of dynamic vehicle motion. The purpose of this research was to offer a secure routing algorithm for selecting a novel CH to reduce the impact of a collision between two cars. The study indicates that the proposed work outperforms the existing scheme in terms of reliability, end-to-end delivery, delivery ratio, and future work of this kind [8]. Liu et al. have presented a method for predicting driving behavior using the Navie Bayes classifier and several unrelated variables that he calls NBP. In this case, driving behavior is predicted using the Naive Bayes classifier. The primary result of this work was the proposal of a machine learning method for forecasting using Navir Bayes, based on an analysis of the driving behaviour of the two opposing groups. Therefore, it concludes that cars can be divided into two groups according to their drivers' attitudes on speeding and passing [9]. Although not directly related to clustering, it highlights the importance of considering dynamic driving patterns in CH selection, which resonates with our study's approach. To select reliable cluster leaders, Oubabas et al. introduced a new method that combines stability and trust characteristics. The key contribution is a description of a new clustering technique that takes into account safety and reliability problems. It underscores the significance of secure CH selection, which aligns with our study's focus on incentivizing stable CH selection. Because CH perform such an important role, the election of a malicious or compromised node as a CH can have far-reaching consequences that pose a danger to the network's security and resiliency. A reliable means of communication and reliable data storage are crucial to safety. The proposed method is broken down into its two main phases, which are outlined in [11]. The one proposed by Kamakshi and Sriram selects the next individuals to get safety warnings by forming stable communities of cars while taking irrelative mobility into account. The fundamental contribution of this work is the realization that clustering algorithms, a method for categorizing people based on their degree of similarity, are widely employed in a variety of online communities. A community detection technique to cluster automobiles with shared characteristics has been implemented. Numerical simulations support the findings that stable communities, reduced packet impact ratios, packet loss ratios, and end-to-end delays, and improved performance and accessibility can be achieved by giving the degree of cohesiveness 30% of the weight and the degree of flexibility 70% of the weight [13].

2.3.2 Spectral clustering based studies

A novel two-pronged strategy for long-term CH selection was proposed by Jabbar and Trabelsi. The first part of the proposed method treats the automotive connection as a one-to-many connection network, which is somewhat close to reality. The work centers on the suggestion that TTM can be used to practically and advantageously break up the hypergraph. This paper provides a thorough analysis of the most crucial considerations for selecting appropriate

clustering parameters, thereby revealing fresh perspectives and new lines of inquiry. A linear timeline offers all the benefits but almost no repercussions. All these internal network nodes serve primarily as communication hubs in this paper. This results in a nearly sparse adjacency matrix, which in turn greatly reduces the overall complexity of the procedure. Thus, it concludes that the number of clusters produced over time affects computing efficiency. When compared to the roads, urban areas offer easier mobility options. CH is also more stable in metropolitan settings, where there are more vehicles per square mile but less freedom of movement [15]. VANET is represented as a hypergraph network clustered using tensor trace maximization with the modularity matrix [29] as the feed in the article [22], which provides a unique modularized hypergraph clustering scheme. Based on four vehicle criteria, including vehicle speed relative to neighbors, consensus trust score, link duration, and connectivity level, CH is selected from each cluster. All parameters are transformed into Eigenvalues before the final CH is chosen. Based on a set of four criteria, the fuzzy decision-making scheme chooses the optimal CH for all vehicles in that cluster. Using cutting-edge techniques, the suggested algorithm for identifying the stable CH for enhanced network performance has also been validated and greatly enhanced. The use of a modularity matrix enhances tensor trace maximization for calculating Eigenvalues, and CH stability increases by 17.64% as compared to no modularity matrix. This motivates the author to create four metrics for vehicles: link longevity, consensus trust score, vehicle connectivity, and relative speed. These vehicle metrics provide a weighted adjacency matrix and a sub-hypergraph for each car that corresponds to an adjacent car. These measurements were improved using the modularity matrix idea, leading to modularized metrics. Fuzzy decision-making is employed to determine which vehicle index within a cluster should be designated as the CH [22].

In order to increase the dependability of vehicular communications, paper [23] proposes a unique cluster-based VANET oriented evolving graph (CVoEG) model by extending the existing VoEG model. Cluster members (CMs) and CH (CHs) are chosen based on network reliability. The suggested CVoEG model separates VANET nodes (vehicles) into an optimal number of clusters (ONC) using the Eigengap heuristic. In terms of packet delivery ratio (PDR), end-to-end delay (E2E), dependability, reliable routing request (RRR), and throughput, the suggested method greatly exceeds the existing schemes. This model proposes a unique approach for cluster formation based on network reliability. Our proposed model extends this by incorporating Modified Tensor Trace Matrix (mTTM) clustering and incentivized CH selection parameters for enhanced network performance. Moreover, for more reliable data distribution and less end-to-end delay in VANET, a novel method called Deep Learning Ruzicka Spectral Clustering-based Oppositional Dragonfly Optimized Routing (DLRSC-ODOR) is introduced by Radhika and Bhuvaneswari in [26]. A deep learning idea is utilised by the DLRSC-ODOR method, which consists of an input layer, two hidden layers, and an output layer. Vehicle nodes are clustered using Ruzicka similarity-based spectral clustering according to several mobility metrics such as vehicle density, movement direction, distance, and velocity. To identify the best possible CH across all criteria, the Multi-attribute Oppositional Dragonfly Optimization method is employed. The output layer then determines the best path between the source and the destination. Results demonstrated that compared to state-of-the-art approaches, DLRSC-ODOR significantly improves reliability, throughput, end-to-end delay, and packet drop rate [26]. This method utilizes deep learning and spectral clustering for efficient routing in VANETs. Our model complements this by introducing Modified Spectral Clustering and incentivized CH selection parameters, enhancing network reliability and performance. A novel approach to recasting the clustering problem as a cutting graph problem is presented by Liu et al. in a paper [27]. Combining the Spectral Clustering algorithm with the enhanced force-directed algorithm, a new, innovative clustering technique has been developed. To improve overall system stability, it uses the mean lifetime of all clusters as an optimization target. The Simulation of Urban Mobility (SUMO) creates several plausible hypothetical situations. The numerical findings show that the method excels at preserving cluster-wide stability. Hypergraph Clustering (HGC), a heuristic clustering based on the hypergraph theory that seeks to optimize the energy of sensor nodes, was proposed in [28] by Gbadouissa et al. theoretical analysis revealed that the clustering process required minimal energy during the cluster formation phase and the determination of the CH. Our proposed model extends this by introducing a dynamic behaviour assessment mechanism for CH selection, ensuring stable and reliable CHs in VANETs.

2.4 Research gap

The examination of various clustering and CH selection methodologies in VANETs can be categorized as spectral clustering-based and non-eigen clustering-based. It also represents a diverse landscape of approaches aimed at improving network performance, reliability, and safety.

- Papers such as [1–3, 6] introduce innovative techniques like the GOA, NSVC, and weighted clustering protocols (non-eigen-based), each addressing high mobility and stability-based challenges within vehicular environments.

- Moreover, research efforts such as those in [8, 9, 11, 13] focus on enhancing network safety and reliability through secure routing methodologies, prediction of driving behaviour, and the selection of reliable CH.
- Most studies highlight the importance of considering factors such as [3, 11, 15, 22], relative speed [15, 22, 23, 27] for selection of CH. However, considering dynamic driving patterns like those shown in [8, 9, 11, 13] in CH selection and clustering algorithms is still need to adopted in CH selection in VANET.
- Furthermore, novel strategies like the spectral clustering-based approach proposed in [15] and the hypergraph-based clustering method described in [22] offer unique perspectives on addressing clustering challenges in VANETs. These methodologies leverage advanced techniques such as tensor trace maximization and modularity matrix calculations to optimize CH selection and cluster formation. However, as the vehicle's density increases, the information stored in adjacency-based matrices becomes less content. Hence, there is a need to improve the spectral clustering-based model

However, despite the advancements made in clustering and CH selection techniques, there remains a research gap in the development of a comprehensive and incentivized CH selection mechanism tailored for vehicular networks. The proposed model here adopts an adaptive CH selection process that incentivizes vehicles based on their adherence to usual driving patterns and behaviours. Hence, the study is motivated by this opportunity for further research to explore the integration of dynamic driving behaviour profiling for reliable CH selection for VANET communication.

2.4.1 Distinguishing features: contrasting our approach with existing models

- Existing methods lack a definitive approach to deciding the optimal number of clusters in VANETs [15, 22, 26], leading to inefficiencies in resource allocation. Here the eigen heuristic approach is employed in the proposed model, which accurately determines the optimal number of clusters with uniform loading, addressing resource allocation challenges effectively.
- Non-spectral clustering approaches struggle with sparse networks and dynamic environments [2–9], limiting their practicality. While hypergraph representations offer promise [15, 22], they often rely on simplistic clustering algorithms like k-means. Here to overcome this, the hypergraph representation using a modified tensor trace maximization approach coupled with GMM clustering instead of k-means is proposed. This allows for probabilistic modelling and better adaptation to the distribution of vehicles in the network, making it more suitable for real-world VANET scenarios.
- While factors like relative speed are commonly considered in CH selection, dynamic driving patterns are often overlooked. Existing models often use a weighted sum approach for CH selection criteria, where weights are arbitrarily assigned by the operator, leading to subjective decision-making. By integrating dynamic driving patterns into our CH selection and clustering algorithms, inspired by the findings of previous studies [8, 9, 11, 13], our model ensures a more adaptive and context-aware approach to CH selection in VANETs.

3 Problem conception

3.1 A Hypergraph approach for system modeling based on modularity matrix

The cluster generation in VANET is not a novel concept but yet to be addressed. Where graph representation is used by many scholars, it is only applicable in sparse areas, where mobility is less. Hence, it is possible to mathematically represent a hypergraph as a set of nodes and hyperedges as shown in Fig. 2. The nodes in a VANET are the cars or other entities involved in communication, whereas the hyperedges are the connections between the nodes as seen in Fig. 2 Let V be the vertex of set of nodes (vehicles and other communication entities) in the hypergraph, and E be the set of hyperedges [22]. In this work the Hypergraph for vehicles is formulated as,

The vertex of a vehicle $V = \{v_1, v_2, v_3, v_4, \ldots\ldots\ldots v_n\}$ of hypergraph the distance metric for another vehicle in the reach is node $E = e_1, e_2, e_3, e_4, \ldots\ldots\ldots e_n$ and w is the weight as mentioned in Table 1. Then the weighted hypergraph is mathematically given as, $H = (E, V, w)$. It is possible to express the connections between nodes in a graph or hypergraph using an adjacency matrix. The adjacency matrix A of a VANET, when represented as a hypergraph, is a $|V|$ x $|V|$ matrix. The element A_{ij} in the matrix is equal to 1 if there is a direct communication link between nodes i and j, and 0 otherwise. The communication links between nodes in a VANET must be defined before an adjacency matrix can be constructed for the network. Consider a VANET with five vehicles, v_1, v_2, v_3, v_4, and v_5. Assuming v_1 and v_2 communicate with one another directly, likewise v_2 and v_3, and v_3 and v_4. Furthermore, v_4 and v_5 vehicles are able to exchange information with one another. Using this information, the adjacency matrix A_{ij} is constructed as follows: $A = [0100010100010100010100010]$. For example, $a_{12} =$

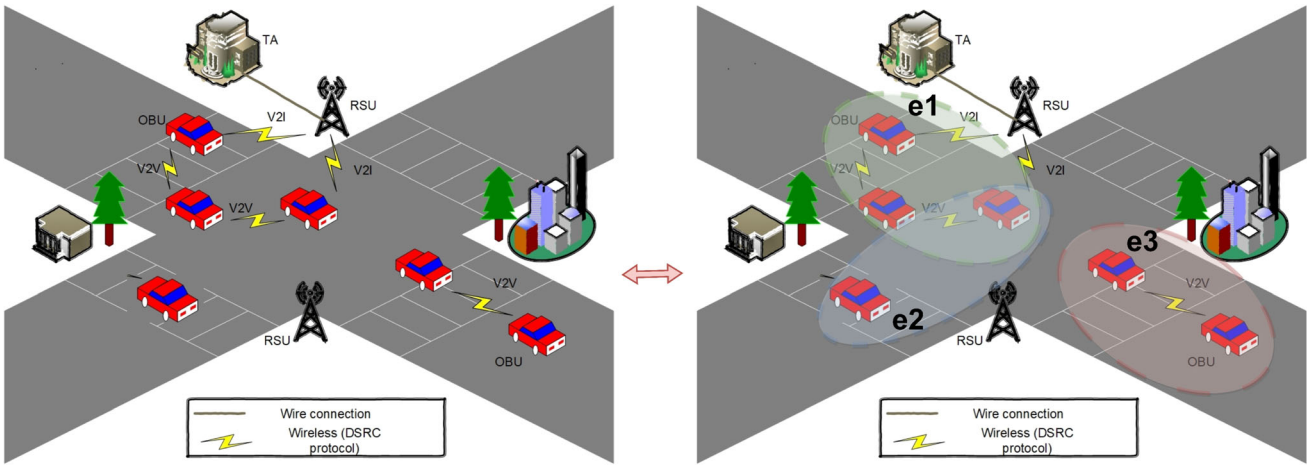


Fig. 2 VANET as hypergraph

1 because there is a direct communication link between nodes v_1 and v_2. Instead of an adjacency matrix, which is typically used to describe a VANET as a hypergraph, the modularity matrix given in [29] is taken into consideration in this research. The modularity matrix is represented as Q. The quality of the partition of the network into communities can be represented by the modularity matrix, which is a $|V|$ x $|V|$ matrix [29] defined as:

$$Q = A - P \tag{1}$$

where A is the adjacency matrix and P is the expected adjacency matrix. The expected adjacency matrix P is defined as in Eq. 2:

$$P_{ij} = \frac{V_{ki} V_{kj}}{2m} \tag{2}$$

where V_{ki} and V_{kj} are the degrees of nodes i and j, respectively, and m is the total number of edges in the network.

The VANET's hypergraph structure can be represented by the modularity matrix Q once it has been defined. Each node in the network is represented by a row and column in Q, and the matrix components indicate the relative importance of the relationship between vertices i and j, taking into account the network's modularity. Most heavily weighted edges are those that connect nearby vertices, which helps in grouping. Removing the weak edge with the lowest $\boldsymbol{Q_{ij}}$ value divides the hypergraph into smaller units. As a result, the modularity matrix provides a more accurate and nuanced picture of the network structure than the adjacency matrix does because it accounts for the predicted number of connections between nodes based on their degree.

Despite the clustering with hypergraph, these clusters show high velocities typically being associated with aggressive driving styles [20]. Given the challenges associated with disruptive behavior caused by impulsive and aggressive drivers, the study proposes an incentive-based CH selection mechanism. This introduces a dynamic behavior estimation process, where nodes (vehicles) are assessed for their behavior within the cluster. The reward mechanism aims to favor nodes exhibiting usual and stable behaviors, increasing their likelihood of being elected as CH. The selection process involves calculating metrics such as Adaptive Neighborhood Degree, Adaptive Link Lifetime, and Adaptive Relative Average Speed. This is further detailed in Next Section.

4 Proposed methodology

The methodology proposed in this study consists of two distinct phases: Modified Spectral Clustering and Stable and Secure CH selection, as illustrated in Fig. 3. The adopted spectral clustering method in this study is detailed as two primary steps: optimal cluster selection explained in Sect. 4.1.1 and mTTM (Modified Tensor Trace Matrix) clustering is explained in Sect. 4.1.2 which utilizes Gaussian mixture models. During Phase one, the simulation commences by incorporating the highest possible vehicle density, thereby collecting data on the vehicles' locations. The hypergraph is constructed based on the provided information and stored in the form of a Laplacian matrix. The modularity matrix is utilized to modify the construction of this matrix. Based on the available data, the Eigen heuristics method is applied to determine the most suitable number of clusters. In the Phase two, the incentivized CH selection process for stable CH. One of the challenges encountered in clustering, particularly in the context of spectral clustering, pertains to the identification of the optimal number of clusters within a VANET simulation. In VANETs, the precise quantity of clusters cannot be

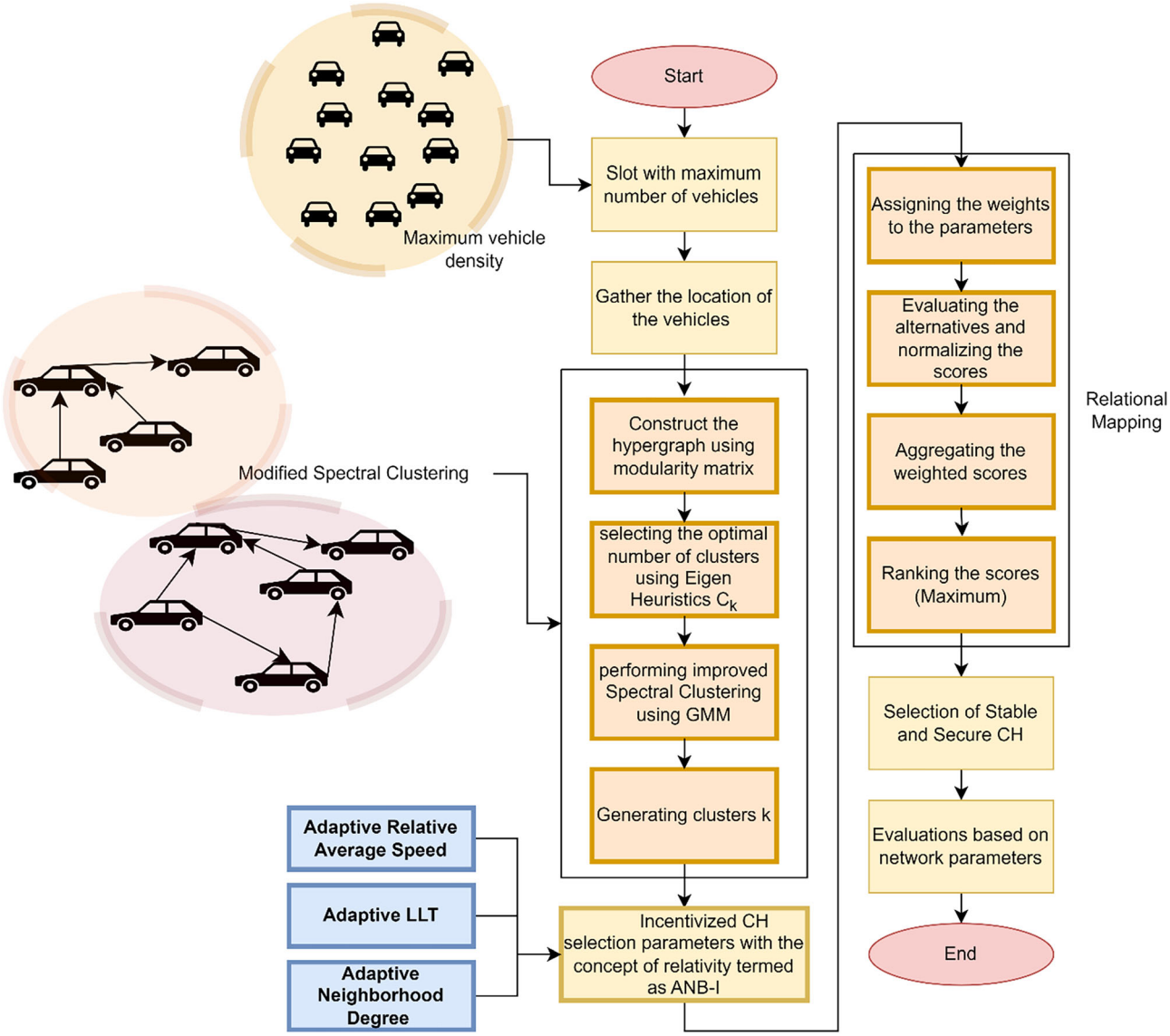


Fig. 3 Detailed proposed framework

predetermined and is subject to fluctuation based on the prevailing network circumstances. Hence, the Eigen Heuristics is employed here for generating an optimal number of clusters.

4.1 Cluster formation

Spectral clustering serves as a fundamental component in the proposed methodology for CH selection in VANETs. Its ability to handle complex data structures and optimize cluster formation contributes to the study's objectives of enhancing network stability, reliability, and responsiveness. Through the incorporation of spectral clustering techniques, the study aims to address the evolving challenges of vehicular communication and cooperative driving in dynamic environments. Here with the help of Eigen heuristics the optimal number of clusters suited for the particular density 100 and 1000 is obtained based on which our mTTM model is implemented for cluster formation in VANET.

4.1.1 Optimized cluster generation

Clustering is an efficient approach to partition the network into smaller, more manageable sections to decrease the excessive burden. A wide range of supervised and unsupervised learning techniques may be found in the literature for classifying networks. For example, K-means [10, 15, 22], machine learning based [9]. The primary limitations of K-means are the requirement to specify the

number of clusters in advance and its inability to accurately categorize vehicles that are not linearly separable in the input space [21]. In the dynamic environment of VANETs, the topology structure undergoes continuous modifications, and it is not guaranteed that the number of clusters will remain constant and K-means is a non-convex optimization problem. In addition to machine learning classifications, other existing clustering techniques [15, 23, 26] employ distinct criteria for cluster creation and CH selection. Determining the optimal number of clusters is a difficult task in all existing clustering methods. Based on the literature [15, 23], having a higher number of clusters can lead to higher hand-off costs, whereas having fewer clusters can result in more frequent link disconnections. The clustering of a developing network relies on the connectedness of nodes at each instant. Therefore, it is preferable to utilize a semi-convex methodology that relies on data connectivity rather than geometrical proximity. In this study, optimal number of clusters is determined using Eigen Heuristics method [23] as shown in Fig. 3. This method analyzes the Eigenvalues of the similarity matrix and identifies the largest gap between consecutive Eigenvalues. The number of clusters is then set to the corresponding number of Eigenvectors.

The Eigen heuristic is calculating the largest difference between successive Laplacian Eigenvalues and then selecting the k value that corresponds to that difference. In this paper, Eigen Heuristics are used to create the best possible value for the Number of clusters. The Eigenvalues of the Laplacian graph L are frequently utilized as input when determining the ideal number of clusters [23]. Input is provided by the Eigenvalues of the Laplacian matrix L, which has a dimension of $n \times n$ (where n is the number of nodes/vehicles). The Eigenvalues of L as $\lambda_1, \lambda_2, \ldots, \lambda_n$ in ascending order, then the Eigen gap heuristic can be used to estimate the number of clusters k as:

$$k = arg\ max(\delta_1, \delta_2, \ldots, \delta_{n-1}) \tag{3}$$

where, $\delta_i = \lambda_{i+1} - \lambda_i$, for i = 1, 2, ..., (n−1) is the difference between successive Eigenvalues, sorted in ascending order. Number of clusters is calculated by using the index returned by the *argmax* operation, which points to the greatest value in the sequence of differences. The Eigen Heuristics method [23] is employed to automatically ascertain the optimal number of clusters without the need for prior knowledge or manual selection. Upon implementation of this methodology, a set of k optimal clusters are obtained, which is subsequently utilized in the application of the modified spectral clustering technique (mTTM). The initial step of this procedure involves the computation of the Laplacian matrix, which is dependent on the interconnections among the vehicles. The use of Eigen Heuristics helps determine the optimal number of clusters, providing a foundation for efficient cluster formation. The spectral clustering method has been enhanced by incorporating the concept of modularity during the computation of the Laplacian matrix. These matrices ensure that the hypergraph captures its essential structure. Furthermore, in this procedure, we have implemented the utilization of GMM as an alternative to k-means for the purpose of assigning vehicles to clusters. The GMM is a probabilistic clustering method that represents the data distribution as a combination of Gaussian components. This approach enables the modeling of diverse cluster shapes and the accommodation of different cluster densities. In this context, k is employed to initiate the clustering process. After the optimal number of cluster is formed, the mTTM clustering is initialized.

4.1.2 mTTM clustering

Spectral clustering involves projecting the data into a lower-dimensional space using the eigenvectors of a matrix. Spectral clustering is a popular unsupervised clustering method used in a variety of fields, including VANETs. In this study, firstly, by traversing the hypergraph that the spectral clustering matrix was based on, a diagonal matrix is generated, it is denoted by *Dig*, the arithmetic sum of which runs over all the vehicle nodes that are next to the node *vi* [22].

$$Dig_{ij} = \sum_{j=1}^{N} Q_{ij} \tag{4}$$

The next step is to calculate the Laplacian matrix L, here the modularity matrix Q is introduced it is given as in Eq. 5.

$$L = Dig^{-1/2} Q Dig^{-1/2} \tag{5}$$

This defines the reminiscent of the clusters that emerge in the simulated road system. To obtain the best cluster configuration for VANET upkeep, these divisions are punned considerably as per the Fig. 3. The modularization of spectral grouping for hyper graphic vehicles is presented in Algorithm 1.

Algorithm 1 Hypergraph based modified TTM clustering with modularity approach

Input: Slot with maximum vehicles: $Vehi_{max}$, Location of each vehicle

1. Select the time slot t when maximum vehicles are present ($Vehi_{max}$)
2. Form a hypergraph $H = (E,V, w)$
3. Calculate the similarity/adjacency matrix based on the distance proximity between the vehicles
4. Calculate the Modularity matrix using Equation 3.1
5. Calculate a diagonal matrix Dig using Equation 4.2
6. Calculate the Laplacian matrix L using Equation 4.3
7. Finding the optimal number of clusters k using Equation 4.1 [23].
8. **For** $i = 1{:}k$:
   a. Cluster assignment for each row is computed using GMM model using $p(x) = \sum_{i=1}^{k} \delta_i N(\mu_i,\Sigma_i)$ where k is number of clusters, δ_i is the probability for point to be in cluster i. $N(\mu_i,\Sigma_i)$ is the normal distribution of mean μ_i and covariance matrix Σ_i. [16].
   b. The clusters $C = \{C_1,C_2,\ldots,C_k,\}$ are generated
9. End For

Output: Cluster Generation $C = \{C_1,C_2,\ldots,C_k,\}$

4.2 Incentivized stable CH selection parameters

In this study, an incentive-based CH selection mechanism is introduced tailored for vehicular networks. A key aspect of this selection is the incorporation of dynamic behaviour of driver's estimation based on the concept of incentive process. Nodes (Vehicles) are assessed for their dynamic behaviour within the cluster, distinguishing between usual and unusual patterns of velocity. The reward mechanism is designed to acknowledge and favour nodes exhibiting usual behaviours, or patterns thereby increasing their likelihood of being elected as CH as elaborated in Fig. 4.

To achieve this, in this study, vehicles initiate the CH selection process by calculating the metrics, Adaptive Neighborhood Degree, Adaptive Link Lifetime (Adaptive-LLT), and Adaptive Relative Average Speed as explained in Fig. 4. The main focus is adopting the concept of dynamic driving patterns and behaviour profiling within vehicular clusters as seen in [20], providing insights into the intricate dynamics and potential driving style switches across different contexts. The authors discuss that high variance in the velocity indicates an aggressive driving style resulting in unusual behaviour and these vehicles will disrupt the communication in clusters making it unreliable. The selection algorithm prioritizes vehicles for CH election based on their Adaptive Relative Average Speed within the cluster and is incentivized.

4.2.1 ANB-I based process for CH selection

To encourage the selection of stable CH, in this study a unique incentivization process incorporating a reward and penalty mechanism based on the dynamic behavior of vehicles termed as ANB-I CH selection parameters. This dynamic behaviour of vehicles aims to align the metrics of neighborhood degree (Adaptive-Nd), Adaptive Relative Average Speed (ADAPTIVE-RAS) and Adaptive-LLT characteristics of vehicles, ensuring a stable and reliable CH.

These metrics ($CH_{metrics\,i}$ for (i = 1 to 3)) are subjected to the relativity and incentive process through the conditions,

$$Adjus_{CH\,metrics\,i} = \begin{cases} Ini_{CH\,metrics\,i} + \delta, (if\ \ diff \leq Thr) \\ Ini_{CH\,metrics\,i} - \delta, (if\ \ diff > Thr) \end{cases} \quad (6)$$

where $\text{Adjus}_{\text{CH}_{\text{metricsi}}}$ is the adjusted CH metric i, $Ini_{CH\,metrics\,i}$ is the initial value of the metric i, δ is the small constant and Thr is threshold value subjected to the difference between the expected and average value of the metric ($diff$). Based on this, the vehicles in clusters at time stamp (t) are rewarded with an incentive or penalized. Adaptive-Nd is defined as the number of corresponding real neighbors of a vehicle whose speed differences are lower than a certain threshold (thr). The RSA of a vehicle is modeled as a random variable following a normal distribution with mean and variance. Similarly, the speed difference as well, so, to determine the threshold is subjected to a dynamic parameter based in standard deviation. It is given as, $Thr = b\sigma$. The description of the metrics for CH selection is given as,

a. Adaptive-RAS: To promote non-aggressive driving and stable speed among the vehicles, an incentivization mechanism for adaptive-RAS is introduced [17]. This process ensures that the vehicles maintain a position close to the average and are rewarded with a chance to be a potential CH. Here, the current speed of the vehicle (X_v) and the average speed of vehicles (X_{avg}) are calculated at the time interval t Whereas in this case, $diff = X_v - X_{avg}$ is subjected in Eq. 3. where the constant is initially set as 0.01 with Adaptive-RAS as 1. This Adaptive mechanism maintains the speed over varying vehicle density compared relative speed. This is shown in Fig. 5.

Compared to the old relative speed parameter for the varying vehicle density the Adaptive-RAS shows less

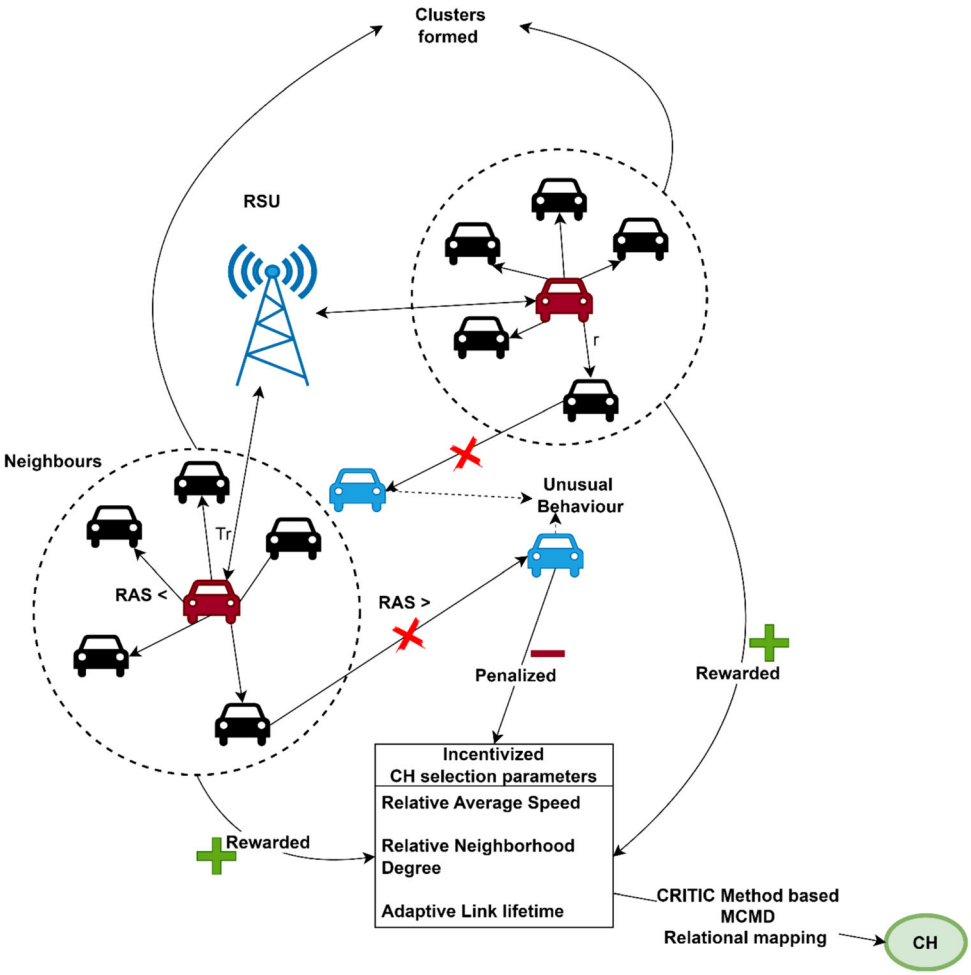


Fig. 4 Incentivized CH selection process

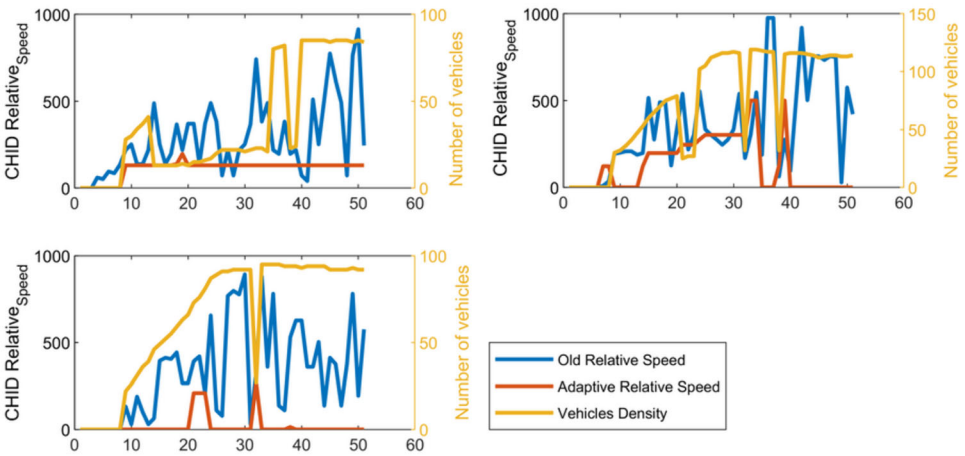


Fig. 5 Evaluation of the proposed incentivized Adaptive-RAS parameter

variation. In the broader context of the driving behavior, where high variance in velocity is associated with aggressive driving styles and potential disruptions in communication within clusters [20], the proposed Adaptive-RAS shows consistency over varying vehicle density contributing towards a stable and reliable selection of CH.

a. Adaptive Nd: During the initial clustering, the vehicles determine the neighbours within the transmission range and the Adaptive-Nd. However, to determine a reliable neighbour, the vehicles showing lower speed variance than the threshold (*Thr*) are rewarded (Eq. 3) and counted as a real neighbour of the vehicle for the given timestamp (t).

$$\begin{cases} RNd(t+1) = RNd(t) + \delta \times 1, \ \ if \ \ diff < thr \\ else, 0 \end{cases} \tag{7}$$

where the neighborhood degree is updated based on the reward mechanism through Eq. 5 and subjected to relativity from Eq. 3, promoting the consideration of vehicles with lower speed variance as reliable neighbors in the clustering process. The proposed method is also evaluated for old Nd and Adaptive-Nd in Fig. 6.

The stability of a CH is contingent on the stability and reliability of its constituent CM. In the CH selection process, the algorithm prioritizes vehicles for CH election based on their relative average speed within the cluster and is incentivized. By incorporating Adaptive-Nd, which rewards vehicles with less speed variance, the selection mechanism further emphasizes the importance of stability and reliability in the chosen CH. This is evident through the simulation results in Fig. 6, that there is less variance with the parameter over the varying vehicles density. This means a consistent and reliable set of neighboring vehicles for each node (vehicle) within the cluster is formed.

b. Adaptive LLT: It ensures a longer duration of link among the vehicles. Consider the vehicles A at position a_i and a_j with speeds v_i and v_j then to sustain communication link it must follow the condition, $-Tr \leq a_i - a_j \leq Tr$ (Tr is communication range). Hence, the LLT_{ij} is given as the ratio of differences in velocities ($D_v = v_i - v_j$) and difference in spatial coordinates ($D_a = a_i - a_j$). In this study, the Adaptive-LLT is given as,

$$\sqrt{(D_a)^2 + (D_v)^2} \cdot \sqrt{(D_v)^2 + \delta} + \frac{Tr}{D_v} \tag{8}$$

The average link lifetime and the individual link lifetime is calculated for the vehicles and the adjustment of the threshold, a small constant δ is added to the standard deviation $\delta\sigma$ to avoid high variation in the number of neighbours as subjected to Eq. 3. If the relative speed is zero, indicating no explicit information, a predefined speed of 40 is considered. This value is based on the predefined speed in the absence of specific relative speed data. The introduction of a constant allows the Adaptive link lifetime metric to handle scenarios with zero relative speed. This adaptive nature in the proposed model is simulated and the evaluation is shown in Fig. 7. It indicates that vehicles within the cluster are maintaining relatively stable links over time without significant fluctuations. Hence, less variation in these metrics aligns with the goal of stability, making vehicles with such characteristics more likely to receive incentives and be elected as CH.

After these metrics are calculated the output is stored as a matrix Y, $Y = [Adaptive - RAS, Adaptive - LLT, Adaptive - Nd]$. K-means is employed with one cluster and the data point in Y, closest to centroid is selected. Then the critic weights are calculated for selection of stable and reliable CH.

4.3 MCMD decision-making process

The weighted approach for the MCMD (Multi-Criteria Decision-Making) method involves assigning weights to the metrics and combining them to make a decision [30]. In this study, metrics stored in Y are normalized and the Pearson correlation is calculated among the normalized metrics. The amount of variation of metrics (σ) is computed for the metrics as well. Further, for each metric in Y, a conflict measure r is calculated by summing the complement of the correlation coefficients. This measure captures the conflict between the jth process indicator and other process indicators. Mathematically, the weights w_i for each of the metric $CH_{metrici}$ in Y is given as,

$$w_i = \frac{\sigma_{CH_{metrici}} \times r_{CH_{metrici}}}{\sum_{j=1}^{n} \sigma_j \times r_j} \tag{9}$$

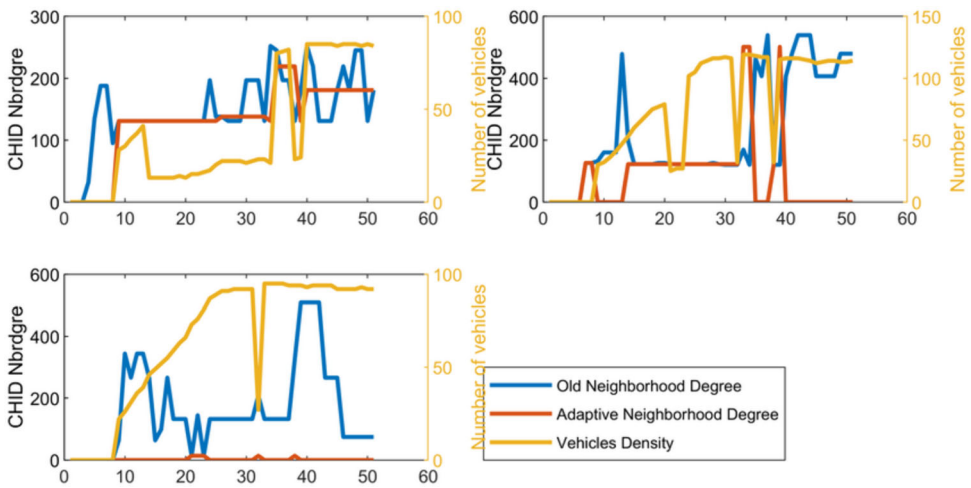


Fig. 6 Evaluation of the proposed incentivized Adaptive-Nd parameter

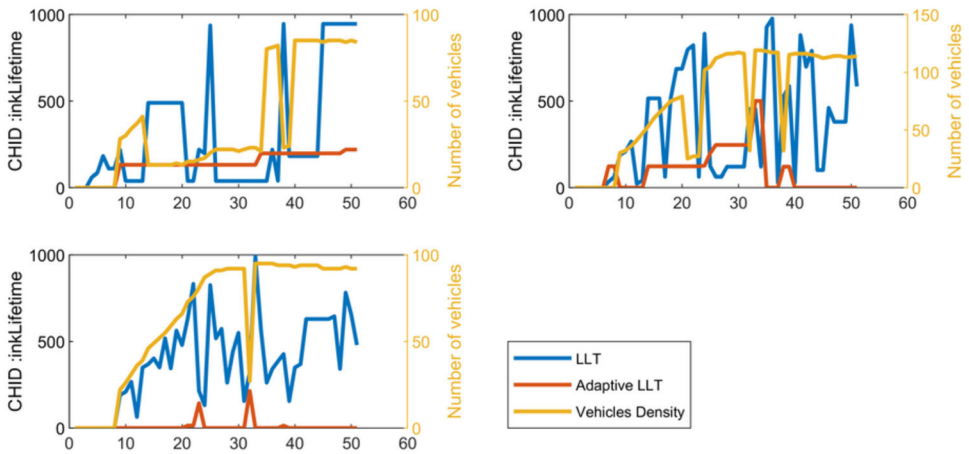


Fig. 7 Evaluation of the proposed incentivized Adaptive-LLT parameter

The final decision is made by considering the sum of metrics with weights multiplied and is given as,

$$W(V_{ID}) = \sum_{i=1}^{CH_{metric\,i}} w_i \times Norm(V_{ID_i}) \tag{10}$$

where, $Norm(V_{ID_i})$ is the normalized value of metric i for the vehicle (V_{ID}). The vehicle with highest weighted sum is considered as CH. It is given as,

$$CH = Max(W(V_{ID})) \tag{11}$$

In this study, weighted approach aims to give more influence to metrics that are both diverse (higher standard deviation) and less correlated with others (higher conflict measure). This aims for a balanced consideration of multiple metrics in the CH selection process. This proposed ANB-I based CH stability improvement model shows consistency over varying vehicle density compared to the individual metrics as seen in Fig. 8. As the process introduces an incentive mechanism that rewards or penalizes vehicles based on their adherence to desirable behavior patterns. By incentivizing vehicles exhibiting usual and stable behaviors, the proposed process encourages the selection of CHs that contribute to a stable and reliable vehicular cluster. This relativity adjustment formulated in Eq. 6, ensures that the CH selection parameters align with real-time behavior, addressing the challenges posed by diverse and dynamic driving patterns observed in different scenarios. Unlike traditional CH metrics that focus on the single aspect of CH metrics, the weighted approach ensures a balanced consideration of diverse metrics, giving more influence to those that are both diverse (higher σ) and less correlated with others (higher r). This leads to a more consistent and balanced decision-making process.

Hence, based on the proposed methodology the simulation is conducted on two varying densities of vehicles 100 and 1000. The discussions and results are given in next Section.

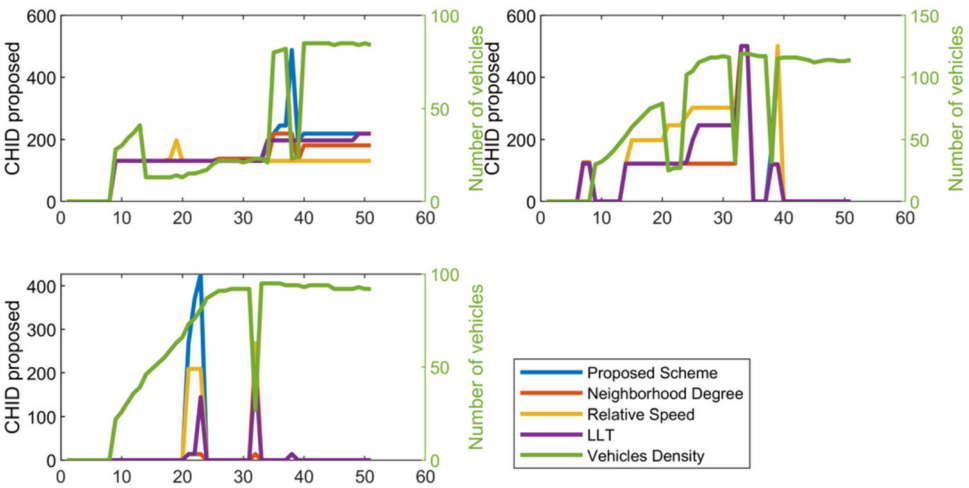


Fig. 8 Evaluation of the proposed incentivized ANB-I model compared to individual metrics

5 Experiments and results

5.1 Simulation parameters

In this Section elaborates the background of the tools employed in this study. The results discussion is considered based on the 2 parts: 1. Network performance (Thr, PD) of the proposed model: this includes the discussion in correlation with CH stability and comparison with state-of-art models. 2. CH stability analysis: this includes the results over packet delay, total distance, throughput and CM frequency over varying vehicle densities.

5.1.1 Tools

The simulation of the proposed model is carried out using the MATLAB (R2023b), SUMO. The complete network model is designed and performance is evaluated through metrics like the throughput and Packet Delay in MATLAB. For the study, the area considered is at 28°39′39″N and 77°13′48″E location in Delhi. SUMO uses Openstreetmap roadways, which have 293 lanes and an average vehicle speed of 10.53 m/s and a maximum speed of 25 m/s, to simulate the traffic conditions at Cannaught Place, Delhi. Delhi Capital of India is a city with huge traffic and is appropriate for an urban setting. The SUMO models traffic flow by utilizing the XML collection of roads and intersections from Openstreet map. The SUMO map of the city considered in this study is shown in Fig. 9.

Three Optimal numbers of clusters are suggested through the proposed Clustering algorithm as seen in Algorithm 1, the clusters in the city are generated and the CH metrics in the matrix Y are recorded for varying vehicle density, i.e., 100 and 1000 vehicles. Through the proposed model 3 clusters were formed and Fig. 10 shows it. Three different colors indicate the clusters and the triangles represent the RSU placed. The simulation parameters used in this study is given by Table 1.

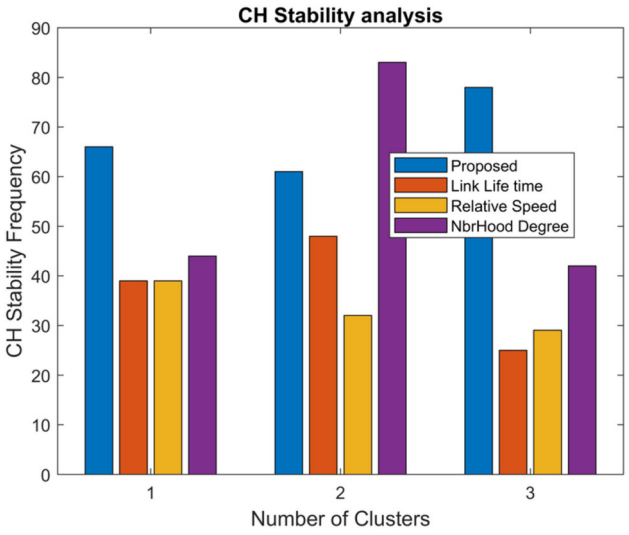


Fig. 10 Cluster-wise evaluation of CH stability for 100 vehicles: Proposed model over individual incentivized CH selection metrics

The communication between cars is achieved through the use of OFDM modulation and the Rayleigh fading channel. The inclusion of the Doppler Effect was necessary to explain the phenomenon of signal degradation resulting from the velocity of a vehicle. The network parameters are obtained from reference [29]. The wireless standard utilized in this study is IEEE 802.11p and the related parameters are noted in Table 2. The standard provides the basis for exchanging data between moving vehicles, mobile nodes, and base stations. Channels with a bandwidth of 10 MHz can be found within the frequency range of 5.9 GHz (specifically, between 5.850 GHz and 5.925 GHz). Due to this progress, receivers will have improved capabilities to handle signal echoes that bounce

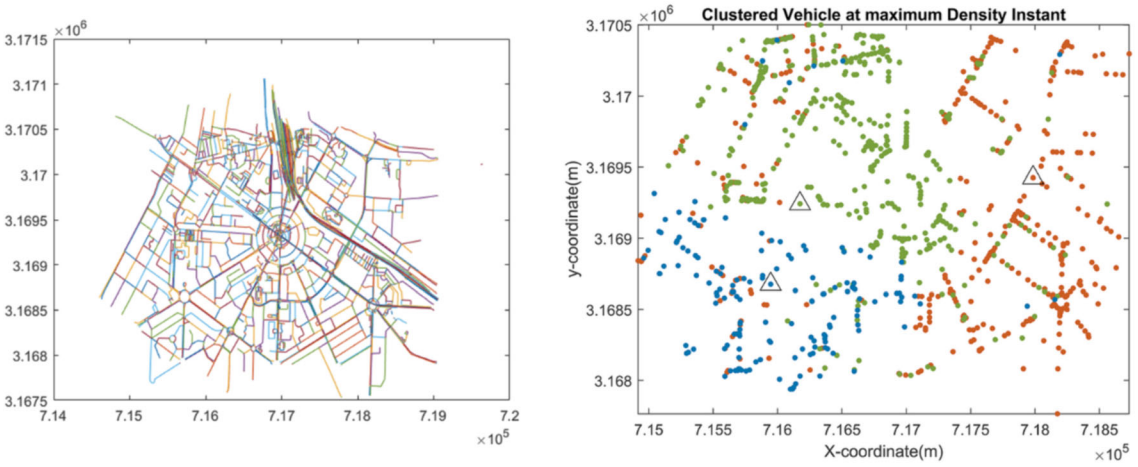


Fig. 9 Map of Delhi and clusters formed at the maximum vehicle density

Table 1 Simulation parameters

Parameter	Value
Density of the vehicle	100, 1000
Lanes	293
Speed of the vehicles	25 m/s
Simulation time	1500 s

Table 2 Specifications of IEEE 802.11p

Attributes	Measure
Doppler shift	$V_{s,r}/\lambda$
Rate of samples	$10MHz$
Number of pilot subcarriers	4
fft point count	32
Subcarrier count	$52 + D$
Data subcarrier count	52
Coding scheme	$1/2$
Modulation Scheme	$OFDM$
Rate of data	$14Mbps$
Cyclic Prefix	16

off other vehicles or buildings in vehicular communication scenarios.

Message dissemination: The process begins with the generation of random bits representing the data to be transferred between vehicles at a rate of 14 Mbps. These bits undergo subcarrier modulation and FFT analysis using 32 FFT points, preparing them for transmission. A Doppler shift is then applied to account for frequency variations due to relative vehicle velocities. Binary Phase Shift Keying (QPSK) modulation is employed to encode the signal onto the subcarriers. Subsequently, white channel noise and Rayleigh channel noise are added to simulate real-world channel conditions. This comprehensive mechanism ensures the effective encoding, modulation, and transmission of data between vehicles within the VANET, considering the specified network parameters and channel characteristics.Top of Form.

5.2 Evaluation parameters

Compared to clustering in MANET where energy and processing power are the main concerns, in case of VANETs its abundant. But, due to high mobility in VANET cluster stability and CH stability are aimed to be improved by many studies. Thus, most of the studies aim to evaluate the clustering stability by considering optimal clusters and improving the network parameters. In most of the studies, cluster stability is defined and evaluated based on CH stability, cluster member frequency, number of clusters and network parameters, but not limited to these. CH stability is evaluated over frequency of change in CM, total distance and Network parameters in this study. When the frequency of CM in clusters increases, the proposed clustering algorithm should be able to maintain the CH stability as well [15]. Hence, in this study, the proposed model is evaluated for the improvement in CH stability over the number of optimal number of clusters formed and proposed ANB-I CH selection parameters process. Moreover, the stability of the proposed model is evaluated through metrics: Throughput and packet delay. The details of the metrics are as follows:

Throughput (Thr) and Packet Delay (PD): At t, the successful transfer of packets from source to destination vehicles is termed as throughput and it is measured in kbs/s. Whereas, when the packets don't arrive on time due to hop travel distance, noise and network congestion, PD occurs.

5.3 Results and discussion

The proposed model for clustering Hypergraph based Modified TTM is analyzed on the Delhi's map under varying vehicle density that is with 100 and 1000 vehicles. When compared to the rural areas with less traffic, metrics in Y carries less significance compared to urban cities like Delhi. Therefore, enabling the priority to consider stable and CH with less variance in speed is necessary. This study also evaluates the obtained results over the proposed Incentivized CH selection process ANB-I.

The proposed model, outlined in the detailed framework and illustrated in Fig. 3, employs a combination of Modified Spectral Clustering and Stable and Secure CH selection for vehicular networks. The model addresses challenges associated with CH stability, frequency of CM, packet delay, and throughput. CH stability Analysis.

5.3.1 Analysis for 100 vehicles

In this study, the clustering algorithm is applied and CH are obtained through Algorithm 2. To evaluate the performance of the proposed model for frequency change in CH stability when the total distance between the vehicles, Throughput, Packet Delay and Change in CM frequency. This is because the proposed clustering algorithm is able to organize the vehicles into optimal clusters with reduced proximal distance. Hence, from the analysis it can be seen that the proposed model is able reduce the CM frequency, packet delay and improve the throughput of network by maintaining the CH stability.

The proposed model introduces an incentive-based CH selection mechanism (ANB-I). The CH stability analysis for 100 vehicles is analyzed cluster-wise over the proposed model verses the individual CH metrics as seen in Fig. 10. It

helps determine if nodes exhibiting desired behaviors are being prioritized as CHs and the incentive process contributes to CH stability. In the optimal cluster 3, the proposed model shows highest CH stability compared to induvial metrics. Whereas, the Relative speed metric shows lower CH stability over 3 clusters and neighborhood degree when applied with relativity process as explained in Eq. 3 shows high CH stability in the cluster 2. The dynamic nature of the relativity process, with the threshold calculated based on the cluster's standard deviation, allows for adaptability to different clusters with varying dynamics. This adaptability enhances the model's robustness and effectiveness in achieving CH stability across diverse scenarios despite the MCMD process applied for mapping the multiple metrics.

5.3.2 Analysis for 1000 vehicles

Lower packet delay signifies improved communication efficiency and faster data transmission between vehicles. The observed outcomes align with the objectives of the proposed model, demonstrating its scalability and adaptability to larger vehicular networks. Moreover, the concurrent improvements in packet delay and CH stability underline its positive impact on network performance. The proposed model shows promising results compared to individual metrics in high vehicle density over cluster 1 as well as 2. This is seen in results obtained from Fig. 11 when the proposed model is analyzed for 1000 vehicles. It is crucial to consider that the proposed model introduces an incentive-based mechanism, prioritizing nodes with usual behaviors and this is why in cluster 3, proposed Adaptive-Nd shows high CH stability compared to the proposed model. Despite the observed discrepancy in cluster 3, the proposed model's emphasis on incentivized stable CH selection remains beneficial for vehicular networks. Hence, the incentive and penalty mechanism align with the study's focus on distinguishing between usual and unusual driving patterns, promoting the election of stable CHs.

Despite a slight discrepancy in cluster 3 for 1000 vehicles, as shown in Fig. 11, the incentive-based CH selection mechanism contributes to overall CH stability. The proposed ANB-I mechanism, prioritizing nodes with desired behaviors, aligns with the model's focus on stable CH selection. The initial evaluation with 100 vehicles demonstrates the proposed model's capability to effectively manage, packet delay, and CH stability. The clustering algorithm, outlined in Algorithm 4.2, plays a pivotal role in organizing vehicles into optimal clusters, mitigating network congestion, and reducing proximal distances. The adaptive nature of the proposed clustering algorithm is highlighted, allowing for efficient organization even with increased vehicle density (1000 vehicles).

5.3.3 Network parameters evaluation

After the CH stability analysis, the proposed model is evaluated for throughput and packet delay for varying vehicle densities 100 and 1000. When 100 vehicles are subjected for simulation on the proposed model, 3 optimal number of clusters are obtained with varying characteristics. The throughput of the proposed model over 2 and 3 cluster is maintained high compared to the individual metrics employed to the simulation as seen in Fig. 12. However, the Adaptive link lifetime adopted in the simulation shows high throughput in cluster 1.

Despite a slight discrepancy in cluster 3 for 1000 vehicles, as shown in Fig. 11, the incentive-based CH selection mechanism contributes to overall CH stability. The

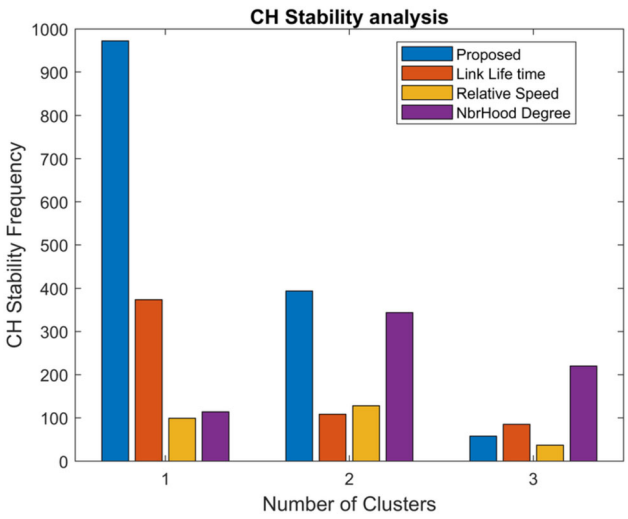


Fig. 11 Cluster-wise evaluation of CH stability for 1000 vehicles: Proposed model over individual incentivized CH selection metrics

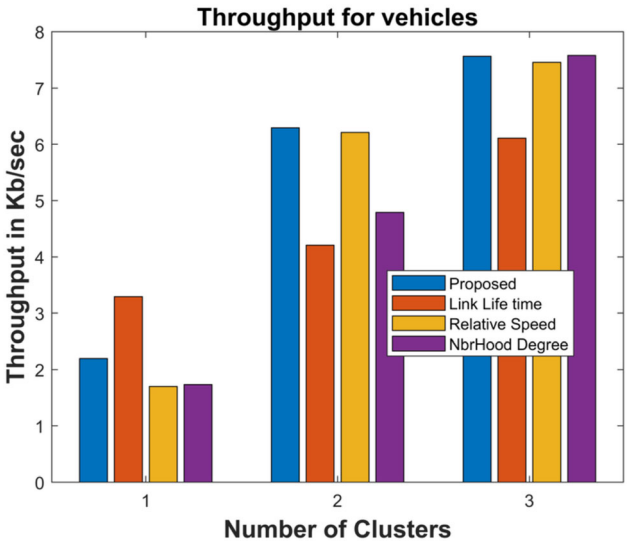


Fig. 12 Cluster-wise evaluation of throughput for 100 vehicles

proposed ANB-I mechanism, prioritizing nodes with desired behaviors, aligns with the model's focus on stable CH selection. The initial evaluation with 100 vehicles demonstrates the proposed model's capability to effectively manage, packet delay, and CH stability. The clustering algorithm, outlined in Algorithm 4.2, plays a pivotal role in organizing vehicles into optimal clusters, mitigating network congestion, and reducing proximal distances. The adaptive nature of the proposed clustering algorithm is highlighted, allowing for efficient organization even with increased vehicle density (1000 vehicles).

5.3.4 Network parameters evaluation

After the CH stability analysis, the proposed model is evaluated for throughput and packet delay for varying vehicle densities 100 and 1000. When 100 vehicles are subjected for simulation on the proposed model, 3 optimal number of clusters are obtained with varying characteristics. The throughput of the proposed model over 2 and 3 cluster is maintained high compared to the individual metrics employed to the simulation as seen in Fig. 12. However, the Adaptive link lifetime adopted in the simulation shows high throughput in cluster 1.

The proposed model's emphasis on incentivized CH selection (ANB-I) introduces a dynamic aspect to cluster formation. The reward and penalty mechanism, based on the dynamic behavior of vehicles, aims to align metrics for stable CH selection as seen in the above discussions in Sect. 5.3.1. This process interacts differently with the CH metrics in clusters, leading to the observed variation in throughput. Whereas, in the case where 1000 vehicles are simulated seen in Fig. 12, shows that the proposed model shows high throughput in all the three cases. This indicates that, at higher vehicle densities, the incentivized CH selection process contributes consistently to throughput enhancement. The dynamic interaction between the reward and penalty mechanism and CH metrics proves effective in diverse scenarios, showcasing the model's potential for enhanced communication stability and efficiency.

It is crucial to have real-time communication in vehicle networks, particularly for safety applications. Timely delivery of vital safety information, such collision alerts, is guaranteed via low packet delay. Congestion and communication issues are more likely to arise when the number of vehicles on the road increases. To determine how well the suggested model scalable to bigger vehicular networks, packet latency and throughput under different densities shows the capability of the proposed clustering and CH selection method.

In case of 100 vehicles for simulation seen in Fig. 13, the proposed model shows low packet delay and Adaptive LLT metric shows highest PD even though high throughput is seen in the Fig. 12. whereas, in the case of 1000 vehicles as seen in Fig. 14, the proposed model shows low packet delay in all the three clusters. In contrast to the variation observed in Fig. 13, the proposed model consistently maintains low packet delay across all three clusters when simulating 1000 vehicles (Fig. 15).

The proposed clustering model, anchored in Hypergraph Modified TTM and featuring an incentivized CH selection mechanism (ANB-I), undergoes a meticulous examination across various vehicular network scenarios, providing valuable insights into its efficacy. In a scenario involving 100 vehicles, the model demonstrates commendable performance in mitigating network congestion, reducing CM frequency, and enhancing CH stability. The clustering algorithm efficiently organizes vehicles into optimal clusters, minimizing proximal distances and fostering improved throughput. A nuanced examination of individual CH metrics reveals dynamic adaptability across clusters, particularly showcased in the cluster-wise analysis (Fig. 10). The incentive-based CH selection process, as seen in Fig. 11, contributes substantially to CH stability, prioritizing vehicles.

The incentivized CH selection process introduces a dynamic aspect to CH selection, leading to varied throughput and PD characteristics over the clusters as well during low mobility simulation (100 vehicles). In contrast, the simulation involving 1000 vehicles (Fig. 14) illustrates the contribution of the incentivized CH selection process to high throughput across all clusters over high mobility. The interplay between the reward and penalty mechanism and CH metrics proves effective, underscoring the model's potential for enhanced communication stability and efficiency. For 100 vehicles (Fig. 13), the model exhibits low packet delay and high throughput, suggesting effective

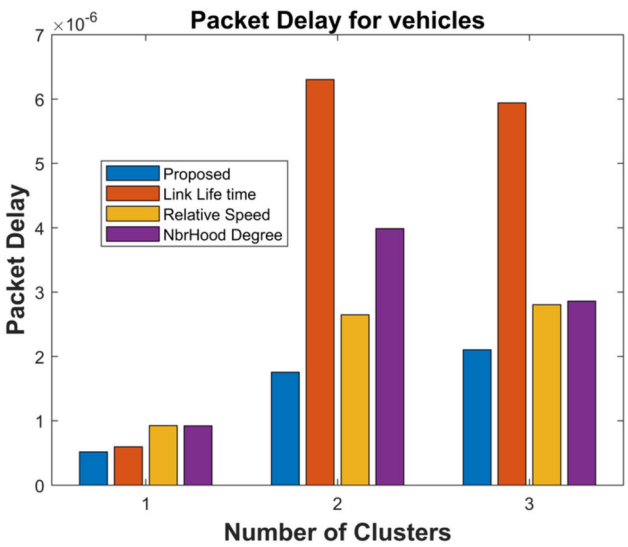


Fig. 13 Cluster-wise evaluation of packet delay for 100 vehicles

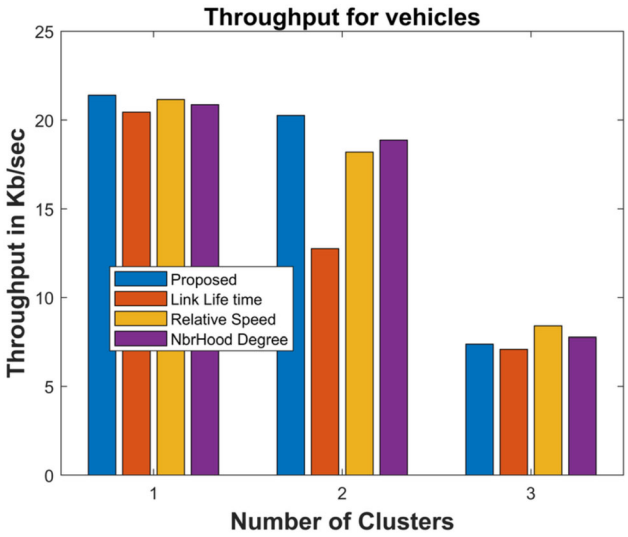


Fig. 14 Cluster-wise evaluation of throughput for 1000 vehicles

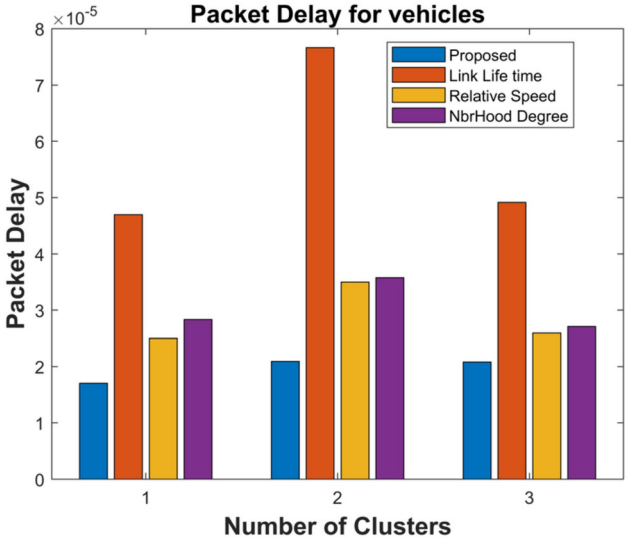


Fig. 15 Cluster-wise evaluation of throughput for 1000 vehicle

communication stability. Notably, the Adaptive Link Lifetime (Adaptive LLT) metric, while contributing to high throughput, introduces higher packet delays in Cluster 1, highlighting that there is a trade-off and optimized network performance for varying vehicle density. However, in the 1000-vehicle scenario (Fig. 14), the proposed model consistently maintains low packet delay across all clusters, showcasing its scalability and efficiency. The proposed model's adaptability to diverse clusters, effective incentivized CH selection, and robust performance in high-density scenarios position it as a promising solution for real-world vehicular communication challenges.

5.3.5 State-of-the art comparison

In this Section, the proposed model is compared with the state of art methods in the same field in terms of CH stability and Network performance. The studies [15, 23, 31] are considered for comparison analysis based on CH stability.

The proposed scheme in [15] shows a key contributions and novelties include the use of hypergraph-based spectral clustering, the incorporation of multiple criteria for CH selection, and the integration of deep learning spectrum sensing for trust calculation. Clustering method in [15] share the concept of hypergraph as in present study. However, the number clusters seem to varying for the vehicle density in [15], compared to the present study. Similarly, [23], introduces CVoEG model divides VANET nodes into an optimal number of clusters using the Eigen gap heuristic. Within each cluster, a vehicle is selected as a CH based on its maximum Eigen-centrality score. The key innovation in [23], is the use of link reliability as a criterion for both CMs and CHs selection.

Vehicular communication is highlighted as an essential component of a smart city. However, scalability remains a major challenge in VANETs, where the dynamic and high-speed movement of vehicles complicates communication and clustering. The proposed JCV algorithm in [23], aims to overcome stability issues by incorporating intelligent and dynamic use of mobility parameters.

These parameters include transmission range, moving direction at the next junction, and vehicle density. This analysis reveals that the proposed model exhibits a 4.17% improvement in CH stability compared to HGCM. It is important to note that while the proposed model achieves a higher CH stability than HGCM, the number of clusters is significantly lower as seen in Table 3.

However, when compared to JCV and CVoEG, the proposed model still maintains a competitive CH stability while requiring fewer clusters. The JCV achieves a slightly higher CH stability of 76% but with 16 clusters, and CVoEG, despite using 20 clusters, lags behind with 65.5% CH stability. This suggests that the hypergraph clustering and ANB-I based CH selection contribute 1.33% and 14.5% improvement compared to [23, 31] respectively, positively to the stability of VANET clusters, making the proposed model a noteworthy advancement in the field.

Next, the network parameters are compared to the proposed model in Table 4. VANETs often support real-time applications, such as traffic safety warnings, collision avoidance, and other time-sensitive communication. Low packet delay is critical for the effectiveness of these applications. This study also aims to ensure the proposed model shows improved QoS and Low delay ensures that

Table 3 State-of-art-comparison based on CH stability

Reference	Method used	CH stability (%)	Optimal number of clusters
Proposed model	Hypergraph clustering and ANB-I based CH selection	75	3
[15]	HGCM	72	4
[31]	JCV	76	16
[23]	CVoeG	65.5	20

Table 4 State-of-art-comparison based on network parameters

Reference	Method used	Packet delay (s)
Proposed model	Hypergraph clustering and ANB-I based CH selection	0.2071
[15]	HGCM	0.42
[32]	JCV	0.27
[23]	CVoeG	0.34

messages or data reach their destination promptly, contributing to a better user experience.

The proposed model, employing hypergraph clustering and an ANB-I (ANB-I) based CH selection mechanism, demonstrates notable superiority in terms of packet delay compared to existing models [15, 23, 32]. Hence, the proposed model shows 102. 84% better reduction in PD compared to [15] and 30.31, 64.171% of better reduction compared to [23, 32] respectively.

The incorporation of the ANB-I (ANB-I) mechanism for CH selection introduces a dynamic and incentive-driven approach. This accounts for real-time driving behavior, promoting the selection of stable CHs. Hence, the proposed model maintains competitive CH stability with fewer clusters compared to JCV and CVoEG. This indicates an effective balance between stability and scalability.

6 Conclusion

The proposed framework introduces a comprehensive solution for clustering and CH selection in VANETs. The model leverages hypergraph-based modified TTM clustering along with an innovative ANB-I approach for CH selection. The hypergraph-based modified TTM clustering algorithm efficiently organizes vehicles into optimal clusters. The algorithm considers the time slot with the maximum number of vehicles, forms a hypergraph, and calculates a similarity matrix based on distance proximity. The modularity matrix and Laplacian matrix are computed to identify the three optimal number of clusters. The study evaluates three main metrics (Adaptive Relative Average Speed, Adaptive Neighborhood Degree, and Adaptive Link Lifetime) for CH selection. These metrics contribute to stable CHs by promoting non-aggressive driving, rewarding vehicles with less speed variance, and ensuring longer link duration among vehicles. he proposed framework is evaluated through simulations using MATLAB and SUMO. The study considers two scenarios with 100 and 1000 vehicles, demonstrating the model's scalability. The results indicate improved CH stability, reduced packet delay, and enhanced throughput compared to individual metrics and state-of-the-art models. The future work of the paper focus on the integration of IoV and Investigate the application of edge computing, edge intelligence in IoV to process data closer to the source. This includes vehicle-to-vehicle (V2V) and vehicle-to-everything (V2X) communication.

Author contributions Ashish Kumari, Shailender Kumar, and Ram Shringar Raw contributed to the study conception and design. Ashish Kumari performed material preparation, data collection, and analysis. All authors reviewed the results and approved the final version of the manuscript.

Data availability Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Ahsan, W., Khan, M.F., Aadil, F., Maqsood, M., Ashraf, S., Nam, Y., Rho, S.: Optimized node clustering in VANETs by using meta-heuristic algorithms. Electronics **9**(3), 394 (2020)
2. Kwon, J.H., Chang, H.S., Shon, T., Jung, J.J., Kim, E.J.: Neighbor stability-based VANET clustering for urban vehicular environments. J. Supercomput.Supercomput. **72**, 161–176 (2016)
3. Khayat, G., Mavromoustakis, C. X., Mastorakis, G., Batalla, J. M., Maalouf, H., & Pallis, E. VANET clustering based on weighted trusted cluster head selection. In *2020 International Wireless Communications and Mobile Computing (IWCMC)* 623–628. IEEE. (2020)
4. Mihelj, J., Kos, A., Sedlar, U.: Implicit aggressive driving detection in social VANET. Procedia Computer Science **129**, 348–352 (2018)
5. Rawashdeh, Z.Y., Mahmud, S.M.: A novel algorithm to form stable clusters in vehicular ad hoc networks on highways. EURASIP J. Wirel. Commun. Netw.Wirel. Commun. Netw. **2012**, 1–13 (2012)
6. Macker, J. P., & Claypool, D. J. Dynamic communities in evolving network graphs. In *MILCOM 2012–2012 IEEE Military Communications Conference* 1–6. IEEE (2012)
7. Zhang, Y., Tie, Y., Liu, Y.: Research on driving behaviour assessment based on evidence theory in 5G-VANET. IET Commun.Commun. **16**(11), 1344–1354 (2022)
8. Iswarya, B., Radha, B.: Stable and secure routing algorithm for VANET. Think India Journal **22**(4), 8115–8120 (2019)

9. Liu, T., Shi, S., Gu, X.: Naive Bayes classifier based driving habit prediction scheme for VANET stable clustering. Mobile Networks and Applications **25**, 1708–1714 (2020)
10. Hu, H., & Lee, M. J. Graph neural network-based clustering enhancement in VANET for cooperative Driving. In 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) 162–167. IEEE (2022)
11. Oubabas, S., Aoudjit, R., Rodrigues, J.J., Talbi, S.: Secure and stable vehicular ad hoc network clustering algorithm based on hybrid mobility similarities and trust management scheme. Vehicular Communications **13**, 128–138 (2018)
12. Kalaivani, D. VANET: Framework, Challenges and Applications.
13. Kamakshi, S., Sriram, V.S.: Modularity based mobility aware community detection algorithm for broadcast storm mitigation in VANETs. Ad Hoc Netw.Netw. **104**, 102161 (2020)
14. Kaur, R., Ramachandran, R.K., Doss, R., Pan, L.: The importance of selecting clustering parameters in VANETs: a survey. Computer Science Review **40**, 100392 (2021)
15. Jabbar, M.K., Trabelsi, H.: A Novelty of Hypergraph Clustering Model (HGCM) for Urban Scenario in VANET. IEEE Access **10**, 66672–66693 (2022)
16. Hu, F., Zhu, Y., Liu, J., Jia, Y.: Computing communities in complex networks using the dirichlet processing gaussian mixture model with spectral clustering. Phys. Lett. A **383**(9), 813–824 (2019)
17. Arkian, H.R., Atani, R.E., Pourkhalili, A., Kamali, S.: Cluster-based traffic information generalization in vehicular ad-hoc networks. Vehicular communications **1**(4), 197–207 (2014)
18. Bando, M., Hasebe, K., Nakayama, A., Shibata, A., Sugiyama, Y.: Dynamical model of traffic congestion and numerical simulation. Phys. Rev. E **51**(2), 1035 (1995)
19. Sharma, R., & Dinkar, S. K. (2022). Selfish node detection by modularized deep nmf autoencoder based incentivized reputation scheme. Cybernetics and Systems, 1–27.
20. de Zepeda, M.V.N., Meng, F., Su, J., Zeng, X.J., Wang, Q.: Dynamic clustering analysis for driving styles identification. Eng. Appl. Artif. Intell.Artif. Intell. **97**, 104096 (2021)
21. Zhou, H., Zhang, D., Xie, K.: Accurate traffic matrix completion based on multi-Gaussian models. Comput. Commun.. Commun. **102**, 165–176 (2017)
22. Kumar, P., Dahiya, P.K., Singh, B.K.: Modularized hypergraph clustering scheme model for stable VANET. Concurr. Comput. Pract. Exp. **34**(26), e7283 (2022)
23. Khan, Z., Fan, P., Fang, S., Abbas, F.: An unsupervised cluster-based VANET-oriented evolving graph (CVoEG) model and associated reliable routing scheme. IEEE Trans. Intell. Transp. Syst.Intell. Transp. Syst. **20**(10), 3844–3859 (2019)
24. Law, M. T., Urtasun, R., & Zemel, R. S. Deep spectral clustering learning. In International conference on machine learning 1985–1994. PMLR (2017)
25. Vashishth, V., Chhabra, A., Sharma, D.K.: GMMR: A Gaussian mixture model based unsupervised machine learning approach for optimal routing in opportunistic IoT networks. Comput. Commun.. Commun. **134**, 138–148 (2019)
26. Radhika, D.: Deep Learned Ruzicka Similaritive Spectral Clustered Oppositional Dragonfly Optimized Routing In VANET. Turkish Journal of Computer and Mathematics Education (TURCOMAT) **12**(12), 3069–3086 (2021)
27. Liu, G., Qi, N., Chen, J., Dong, C., Huang, Z.: Enhancing clustering stability in VANET: A spectral clustering-based approach. China Communications **17**(4), 140–151 (2020)
28. Gbadouissa, J.E.Z., Ari, A.A.A., Titouna, C., Gueroui, A.M., Thiare, O.: HGC: HyperGraph based Clustering scheme for power aware wireless sensor networks. Futur. Gener. Comput. Syst.. Gener. Comput. Syst. **105**, 175–183 (2020)
29. Newman, M.E.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**(23), 8577–8582 (2006)
30. Taherdoost, H., Madanchian, M.: Multi-Criteria decision making (MCDM) methods and concepts. Encyclopedia **3**(1), 77–87 (2023)
31. Mukhtaruzzaman, M., Atiquzzaman, M.: Junction-based stable clustering algorithm for vehicular ad hoc network. Ann. Telecommun.Telecommun. **76**, 777–786 (2021)
32. Kakkasageri, M.S., Manvi, S.S.: Multiagent driven dynamic clustering of vehicles in VANETs. J. Netw. Comput. Appl.Netw. Comput. Appl. **35**(6), 1771–1780 (2012)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Ashish Kumari received her B Tech. from Maharishi Dayanand University, Rohtak, M Tech. from Rajasthan Technical University, Rajasthan, India, and doing PhD. from Delhi Technical University, New Delhi, India. She is working as an Assistant Professor (IT) at the Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, India. She has over 10 years of experience in academia and research. Her research interests include Networking, Machine Learning, and Big Data Analytics.

Shailender Kumar is working as a professor in the Department of Computer Science & Engineering at Delhi Technological University, India. He has more than 23 years of experience in teaching, administration, and research. He has specialization in Database Systems, Data mining, Big Data Analytics, Machine Learning, and Information security. He has published several papers with good impact factors in reputed International Journals and Conferences including IEEE, Elsevier, Springer, Wiley & Sons, etc.

Ram Shringar Raw received his Ph.D. (Computer Science and Technology) from the School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India in 2011. He is currently working as an Associate Professor in the Department of Computer Science and Engineering of Netaji Subhas University and Technology, Delhi, India. He has more than 22 years of teaching, administrative, and research experience. His current research interests include Mobile Ad hoc Networks, Vehicular Ad hoc Networks, Flying Ad-hoc Networks, and IoT Cloud. Dr. Raw has published 14 books, 06 patents, and 140 research papers with good impact factors in reputed International Journals and Conferences including IEEE, Elsevier, Springer, Wiley & Sons, Taylor & Francis, Inderscience, Hindawi, IERI Letters, American Institute of Physics, etc.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19528-7

Modified-generative adversarial networks for imbalance text classification

Poonam Rani[1] · **Om Prakash Verma**[2]

Received: 16 January 2024 / Revised: 19 April 2024 / Accepted: 26 May 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
Textual data faces a significant challenge known as class imbalance. This challenge arises when the distribution of classes within a dataset is highly disproportionate. To address this problem, effective algorithms that can handle the imbalance in the data while maintaining high predictive accuracy are needed. To solve the class imbalance problem, numerous methods have been proposed, which include SMOTE (Synthetic Minority Over Sampling), ADASYN (Adaptive Synthetic Sampling), ensemble methods, etc. In this paper we propose the novel Synthetic Minority Over Sampling Technique Nominal Continuous-Generative Adversarial Network (SMOTE NC-GAN) Model to tackle the problem of imbalanced text datasets. This method provides a novel solution for addressing the class imbalance problem. It involves generating synthetic data from the minority class of dataset using SMOTE-NC and subsequently applying that data along with real data fed to the Generative Adversarial Network. By using a combination of SMOTE-NC and GAN techniques, our proposed method can generate synthetic samples that enhance the quality of minority classes in the text datasets. Our proposed model addresses the limitations of both GAN and SMOTE. Empirical findings on various benchmark datasets have demonstrated the superior performance of the SMOTE NC-GAN Model over different state-of-the-art methods. This model enhances sample quality and offers computational efficiency, making it a promising solution for handling imbalanced datasets across various domains. The classification performance is significantly improved when utilizing the SMOTE NC-GAN Model in comparison to the Synthetic Minority Oversampling Technique.

Keywords Generative adversarial network · Synthetic minority over sampling technique · Natural language processing · Synthetic minority over sampling technique nominal continuous

✉ Poonam Rani
poonam.rani2007@gmail.com

Om Prakash Verma
opverma.dce@gmail.com

[1] Department of Comp Science and Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur, Rohini 110042, Delhi, India

[2] Department of Electronics and Communication Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur, Rohini 110042, Delhi, India

Published online: 12 June 2024

1 Introduction

With the rapid advancements in information and communication technology (ICT), there has been a significant rise in research areas. This progress has provided users with convenient and fast access to information. Consequently, there is an increase in the number of text documents, as the textual form of data is mostly used for storing information [1]. Analysis of data and prediction are commonly used in text classification. Numerous machine learning techniques, including K-Nearest Neighbour [2], Random Forest [3], Support Vector Machine [2], Ensemble classifiers [4], and Naive Baye's [5] are used in various applications of text categorization. Several challenges arise during text classification, and one of the major challenges is class imbalance. The term "class imbalance" denotes the unequal representation of classes in a dataset, i.e., the proportion of instances in one class being much higher than the other class. The issue of class imbalance poses challenges because many machine learning algorithms are designed under the assumption that classes are distributed uniformly within the dataset. However, in reality, this assumption is often violated due to imbalanced class distributions. The model's performance is greatly impacted by the existence of an unbalanced class distribution, which results in biased classification that favours the dominant class. When most samples are categorized as positive, this imbalance leads to low true negative and high true positive rates [6]. Data imbalance is a prevalent occurrence in various scenarios, such as Credit card fraud detection [6], Cyberbullying detection [7, 8] uncommon diseases [9, 10], and the prediction of natural disasters [11]. Cyberbullying datasets face data imbalance problems as the number of normal labeled classes is higher than abnormal labeled ones. In recent years [12], researchers have shifted their focus towards artificial general intelligence (AGI), a field striving to develop software capable of emulating human-like intelligence and possessing the ability to self-learn. AGI aims to empower software to perform tasks beyond its initial training. Progress in brain-inspired intelligence research is crucial for achieving higher intelligence levels, accuracy, robustness and energy efficiency surpassing existing AI systems. Several framework have been proposed for eg.HESFOL [13], MeMEE [14], SNIB [15], SIBoLS [16], HOSIB [17]. Various techniques have been developed to address the issue of imbalanced binary datasets [9]. However, researchers also encounter significant challenges in addressing data imbalances in multi-class datasets [18]. Extended algorithms, designed for binary algorithms, are often the existing solutions for multi-class imbalanced datasets. In multi-class datasets [19], ensemble-based classifiers and decision trees [20] have been shown to yield better results. The problem of class imbalance has received significant attention from researchers. Numerous studies have been conducted to address the problems caused by class inequality [21–24]. In text classification, there are various methods that exist to resolve the class imbalance problem. The classical Synthetic Minority Over-sampling Technique [25] is a widely acknowledged method for resolving class imbalance in text classification. It addresses imbalanced datasets by generating synthetic samples through interpolation between existing minority class samples, focusing only on continuous features. However, the main challenge with SMOTE [26] is how to calculate the distance between points when categorical features are present. The extension of SMOTE, known as SMOTE-NC [26], generates synthetic data for imbalanced datasets with both nominal (categorical) and continuous features. On the other hand, while the primary utilization of GANs (Generative Adversarial Networks) [27] is in generating synthetic images for image processing, GANs have not been extensively applied to small datasets in text processing, especially those representing minority classes. A large amount of data is required to train GAN effectively and only the GAN model can generate meaningful synthetic data. We introduced a novel model named SMOTE_NC-GAN

Model, which effectively tackles the class imbalance problem by integrating SMOTE-NC [26] and GAN [27]. This approach involves generating synthetic data from the minority class of the dataset using SMOTE-NC and subsequently applying that data along with real data to GAN. The performance of our proposed model is evaluated using accuracy, recall, F1-score, and the ROC curve. We performed a comparative analysis of our proposed model with other state-of-the-art models to assess its performance in text classification. The results indicated that our model outperformed the other methods, making it a promising and optimal solution for text classification tasks. The main contribution of this paper is summarized as follows: We proposed a novel approach that combines the GAN (Generative Adversarial Network) and SMOTE-NC techniques to generate synthetic data. Our proposed model offers improved performance and flexibility in handling diverse datasets. We perform the implementation and evaluation of SMOTE and SMOTE NC-GAN, on three imbalanced datasets: Credit Card Default, Adult Income, Cervical Cancer. The performance of the algorithms is compared using evaluation metrics.

The subsequent sections of the paper are structured as follows: State-of-the-art methods to resolve the class imbalance problem are discussed in Section 2. Section 3 presents a detailed description of our proposed method, the SMOTE NC-GAN Model. Section 4 presents the experimental results obtained from our proposed method with other existing methods and analyzes these results. The outcomes of the experiments and their implications are discussed in Section 5. Finally, we discuss the summary and conclusion of the proposed technique and potential areas for future work in Section 7.

2 Related work

In this section, we explored various existing techniques aimed at addressing the class imbalance problem [28]. The first subsection presents a detailed overview of the work related to the imbalance class.

2.1 Class imbalance techniques

The performance of classification algorithms is significantly affected by the Class Imbalance Problem (CIP) due to the introduction of a disproportionate number of instances from one class compared to others [29]. Approaches to address the CIP can be classified into algorithm level [30], data level techniques [31], cost-sensitive learning [30], and ensemble-level approaches [32]. All pre-processing techniques that are aimed at balancing the distribution of classes are categorized into data levels [33]. Sampling methods are used to address the four challenges of class imbalance and can be classified into two prime categories: one is undersampling, and the other one is the Oversampling method [34]. Under-sampling reduces the number of instances from the majority class to achieve a more balanced distribution. However, this approach has the drawback of potential loss of valuable information. Various undersampling techniques have been proposed in the research literature. These methods consist of Wilson's Edited Nearest Neighbor rule, One-Sided Selection [35], Tomek Links [36, 37] and Condensed Nearest Neighbor Rule [38]. These techniques aim to alleviate class imbalance by selectively removing instances from the majority class while preserving the essential characteristics of the minority class. They have gained significant attention and adoption within the research community as effective strategies for addressing class imbalance challenges. Heba Mamdouh Farghaly [39] proposed a model to solve the feature selection problem in

pattern recognition. A new method FS-FAI [40] proposed for the classification of text which is based on associated and frequent itemsets dataset. Esraa Hassan [41] uses the MCN-BERT and BiLSTM models to predict diseases automatically from symptoms. Tarek M Mahmoud [42] introduced an automatic classification system for web pages. In this paper accuracy is enhanced using the combination of K-NN and SVM and Naive Bayesian algorithm. Marwa Khairy [7] addresses the data imbalance problem in the Cyberbullying [43] dataset by examining the efficacy of four resampling techniques which include SMOTE-TOMEK, random oversampling, random undersampling and SMOTE. Marwa Khairy [8] evaluates the efficacy of various single and ensemble machine learning algorithms in detecting Arabic text containing offensive language and cyberbullying by introducing a novel balanced Arabic dataset [44–46] for offensive language detection. Ahmed Omar [47] introduced a multi-label Arabic dataset for multilabel classification tasks, sentiment analysis and short text classification through a combination of semi-supervised and manual annotation approach. Oversampling is a technique in which synthetic or duplicate instances of the minority class are generated to balance imbalanced datasets. However, one of the main risks associated with oversampling is overfitting, as it may create an excessive number of synthetic instances that can overly represent the minority class and lead to poor generalization on unseen data [48]. To address the risk of overfitting, SMOTE (Synthetic Minority Over-sampling Technique) was introduced [26]. To address the different aspects of class imbalance several variants of SMOTE have been developed. These variants include SMOTE [26], and K Means-SMOTE [49], Borderline-SMOTE [31]. G.Kovacs [50] has implemented a comprehensive set of 85 SMOTE variants in a Python library. However, SMOTE is the most popular oversampling to solve data imbalance problems but it also has some drawbacks. SMOTE focuses only on continuous features. However, when categorical features are present in the dataset, calculating distances between points becomes challenging. To address this challenge, an extension of SMOTE called SMOTE-NC (SMOTE for Nominal and Continuous features) has been proposed. SMOTE-NC incorporates a modified distance metric that can handle both continuous and categorical features [26]. Therefore SMOTE-NC is introduced which is an extension of SMOTE that generates synthetic data for imbalance datasets with both nominal (categorical) and continuous features. Similar to SMOTE, SMOTE-NC does have certain drawbacks. These include suboptimal performance when dealing with high-dimensional data [51], the potential to create overfitting and introduce noise into the dataset [51, 52], and the neglect of the local distribution of minority samples [51]. These drawbacks arise due to the interpolation techniques used, the lack of consideration for individual minority samples, and the uniform application of the parameter K neighbors across the entire dataset.

2.2 High dimensionality and class-imbalance

Many oversampling techniques are specifically designed for binary classification problems and assume that the feature space used for classification is low dimensional. However, high dimensional data is used in real-world datasets. High dimensional features in the presence of class imbalance can negatively impact classification performance, leading to inaccurate results [35]. Therefore in high-dimensional space, generating samples using traditional oversampling techniques may not be effective. This limitation highlights the need for the development of oversampling methods that can handle high-dimensional data and effectively tackle class imbalance issues in such scenarios. Blagus [53] conducted a study on high-dimensional class-imbalanced data and found that the Synthetic Minority Over-sampling Technique (SMOTE) is not effective in reducing the classification bias towards the majority

class for most classifiers. The author observed that SMOTE performs even less effectively than random undersampling. The study suggests that applying SMOTE will not effectively address the class imbalance issue in scenarios where the data is high-dimensional and imbalanced. These findings highlight the importance of exploring alternative approaches when dealing with high-dimensional class-imbalanced datasets rather than relying solely on SMOTE to address the class imbalance. This limitation of SMOTE motivates us to explore alternative approaches, such as Generative Adversarial Networks (GANs). Researchers have dedicated efforts to tackle these challenges and enhance the accuracy of classification results. Douzas [54] introduced a generative model called Conditional Generative Adversarial Network to address class imbalance in binary datasets. CGAN utilizes the class labels of imbalanced datasets as conditioning information and generates synthetic data for the minority class. The performance of CGAN was evaluated on various datasets with different characteristics and compared with different oversampling methods using multiple classifiers. The results demonstrated that CGAN outperformed other models in terms of classification performance. They focused on training a conditional GAN network incorporating minority and majority class samples. They aimed to capture the overall data distribution of the imbalanced dataset. In contrast, our approach specifically trained GAN networks using only minority samples. Previous research has shown that Generative Adversarial Networks (GANs) are highly effective in generating synthetic data that closely mimics real data distribution in various domains. GANs have been successfully employed in several applications, including object detection/recognition [55], super-resolution [56], and face aging [57]. GAN is the most popular deep learning model mainly used in the field of image analysis. Giovanni Mariani [58] employed a variant of GAN known as BAGAN (Balancing GAN) as a data augmentation technique to address class imbalance in imbalanced datasets. Their approach involved training GANs to generate synthetic images specifically for the minority class, achieving better convergence and generating high-quality synthetic images. They compared the performance of BEGAN [59] with another generative adversarial network called ACGAN (Auxiliary Classifier GAN) [60]. The results of their study demonstrated that BEGAN outperformed other generative adversarial networks in terms of restoring balance in the imbalanced datasets and improving overall classification performance. This work highlighted the challenges of mode collapse and convergence in GAN training [61, 62]. Berthelot et al. introduced Boundary Equilibrium GANs (BEGANs) [59]. BEGANs incorporated an equilibrium enforcement mechanism to stabilize GAN training and address convergence issues. Karras introduced Progressive GANs (PGANs) [63], implementing a progressive training technique to increase the quality of the sample generated.

3 Theoritical foundations

SMOTE is the most common technique to be used to resolve class imbalance problems. It struggles to create synthetic data that resemble the original data. As a result, researchers have turned to Generative Adversarial Networks (GANs) as an alternative for oversampling, as GANs have been shown to effectively generate synthetic data that closely resembles the original data [64]. Due to the limitations of existing techniques, such as SMOTE, in effectively addressing class imbalance in high-dimensional data, researchers need to explore and develop new approaches. These new approaches should focus on generating synthetic samples that can balance the skewed distribution of high-dimensional data. To overcome this limitation, we combine the Generative Adversarial Networks (GANs) and SMOTE-NC

for oversampling. GANs are effective in generating artificial data that closely resembles the original data distribution. Previous studies [55, 57, 64] have highlighted the potential of GANs in generating realistic synthetic data.

3.1 Generative adversarial network

Goodfellow [27] introduced Generative Adversarial Networks in 2014. It comprises of two-component generator(G) and discriminator(D). The input to the generator is the random noise and this noise is transformed to generate synthetic data. The training of the discriminator is done on real data and it acts as a binary classifier to distinguish between real and fake data. The main goal of the generator is to generate synthetic data that resembles the real data and also minimize the loss function, and the main objective of the discriminator is to find the difference between real data and fake data and to maximize that difference, i.e., loss function. The discriminator minimizes $D(G(z))$ and maximizes $D(x)$ to overall maximize the loss function in (1). The generator and discriminator engage in an adversarial game during the training process. This adversarial relationship between the two networks leads to optimizing both models. This competitive interplay between the discriminator and generator improves both networks over time.

Discriminator loss function is calculated as [27]:

$$L^{(D)} = \max[\log D(x) + log(1 - D(G(z)))] \tag{1}$$

Where D denotes the discriminator that classifies the real data (x) and synthetic data $(G(z))$. $D(x)$ represents the probability distribution obtained from real data (x), while $D(G(z))$ represents the probability distribution of the synthetic data (z). The discriminator aims to maximize $D(x)$ and minimize $D(G(z))$ to maximize the overall loss function given in (1). $D(x)$ and $D(G(z))$ are both probability values ranging between 0 and 1. The generator's aim is to deceive the discriminator by producing synthetic data that closely resembles real ones, which means the generator aims to produce $G(z)$ in a way that, when passed through the discriminator, it receives a classification of 1. The discriminator's goal is to assign a classification of 1 to $D(G(z))$, while the generator attempts to achieve a classification of 0 for $D(G(z))$.

The generator loss function is calculated as follows[38]:

$$L^{(G)} = \min[\log D(x) + log(1 - D(G(z)))] \tag{2}$$

The generator objective is to minimize the overall loss function given in (2) and to minimize the (2), the generator must maximize the $D(G(z))$, and the generator does not have access to the real data $D(x)$.

3.1.1 Combined loss function

Equations (1) and (2) are combined as

$$\min_G \max_D V(D, G) = [\log D(x) + \log(1 - D(G(z)))] \tag{3}$$

The discriminator aims to minimize $D(G(z))$, and the generator aims to maximize $D(G(z))$ given in (3). The validity of this loss function is limited to individual data points. However,

when considering the entire dataset and dealing with the minimization of the discriminator and maximization of the generator in a GAN, the expectation of the equation is computed as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{4}$$

where $P_{\text{data}}(x)$ represents the distribution of real data that matches the data generated by the generator G, i.e., $P_G(x)$. Rather than explicitly assigning probabilities to individual data points, GANs learn to transform a noise variable z drawn from a noise distribution $P_{\text{noise}}(z)$ into a sample $G(z)$ from the generator distribution P_G.

3.2 Proposed SMOTE_NC-GAN model

We introduced a novel model named SMOTE_NC-GAN Model, which effectively tackles the class imbalance problem by integrating SMOTE-NC and GAN. This approach involves generating synthetic data from the minority class of the dataset using SMOTE-NC. This is achieved by considering the local neighborhood of each minority sample and creating new samples that possess similar feature values within that neighborhood. The aim is to augment the minority class so that the dataset gets balanced. The resulting dataset now consists of synthetic samples from SMOTE-NC. The newly generated synthetic data, along with the real data, are fed into the discriminator. The discriminator's main aim is to find the difference between synthetically generated data and real data. The discriminator takes input from SMOTE-NC and GAN. Training the discriminator with a mixed dataset of real and synthetic samples improves its ability to identify and classify the two types of data accurately. Between the discriminator and the generator, there is the feedback loop, which is responsible for producing synthetic samples and encourages the generator to generate more realistic and representative synthetic samples. The discriminator enhances its capability to identify real and fake samples through this adversarial process. The addition of synthetic data to the minority class data makes the minority class equivalent to the majority class, which reduces the model's bias towards the majority class and improves the model's generalization capability, leading to more reliable and robust results.

Our proposed model loss function for the discriminator is calculated as:

$$L^{(D)} = \max[\log D(x) + \log(1 - D(G(z))) + \log(1 - D(S(y)))] \tag{5}$$

where D(x) represents the output of the discriminator when given real data x. The value D(x) is the discriminator's prediction of the probability that the real data is real. The generator takes random input z and produces synthetic data using a mapping function G. G(z) is the generator's output when given noise z and D(G(z)) represents the output of the discriminator when given synthetic data generated by the generator. The value D(G(z)) is the discriminator's prediction of the probability that the synthetic data is real, and S(y) represents the synthetic generated data y from SMOTE-NC. D(S(y)) represents the output of the discriminator when given the generated synthetic data from Smote_NC. The value D(S(y)) is the discriminator's prediction of the probability that the generated synthetic data is real. The discriminator aims to maximize $D(x)$ and minimize $D(G(z))$ and $D(S(y))$ to maximize the overall loss function given in (5). $D(x)$ and $D(G(z))$ and $D(S(y))$ probability values ranging between 0 and 1. The generator loss function is calculated as follows:

$$L^{(G)} = \min[\log D(x) + \log(1 - D(G(z))) + \log(1 - D(S(y)))] \tag{6}$$

The generator aims to minimize the overall loss function given in (6) and to minimize the (6), the generator must maximize the $D(G(z))$ and $D(S(y)$. The generator does not have access to the real data $D(x)$.

3.2.1 Proposed GAN loss function

Equations (5) and (6) are combined as:

$$\min_G \max_D L(D, G) = \left[\log(D(x)) + \log(1 - D(G(z))) + \log(1 - D(S(y)))\right] \tag{7}$$

The discriminator aims to minimize $D(G(z))$, and the generator aims to maximize $D(G(z))$ and $D(S(y))$. This implies that the discriminator's parameters (represented by D) are optimized to maximize the loss function, while the generator's parameters (represented by G) are optimized to minimize the loss function given in (7). The validity of this loss function is limited to individual data points. However, when considering the entire dataset and dealing with the minimization of the discriminator and maximization of the generator in a GAN, the expectation of the equation is computed as:

$$\begin{aligned}\min_G \max_D V(D, G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \\ + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(S(y))]\end{aligned} \tag{8}$$

where $P_{\text{data}}(x)$ is the real distribution of data that matches the data generated by the generator G, i.e., $P_G(x)$. Rather than explicitly assigning probabilities to individual data points, GANs learn to transform a noise variable z drawn from a noise distribution $P_{\text{noise}}(z)$ into a sample $G(z)$ from the generator distribution P_G.

3.3 Architecture of proposed SMOTE_NC-GAN model

The first step in our methodology is the application of Boruta's feature selection algorithm, as shown in Fig. 1. Boruta's algorithm helps identify the most relevant features in the dataset by comparing their importance to randomized versions. Features that consistently exhibit

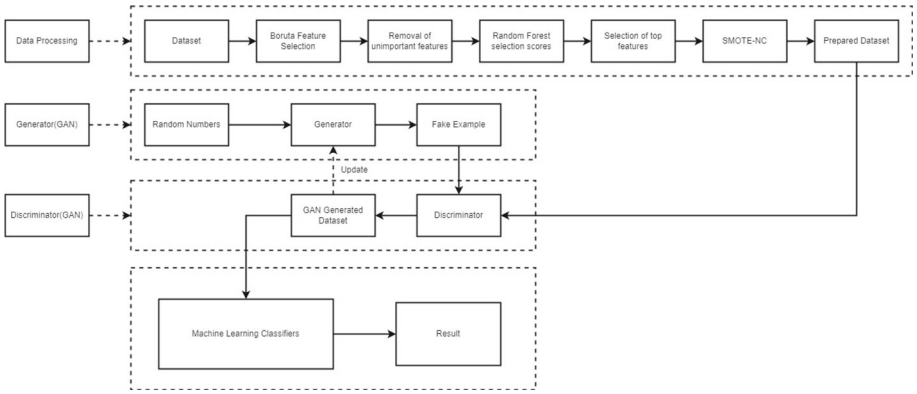


Fig. 1 Architecture of Proposed SMOTE_NC-GAN Model

higher importance than the randomized versions are considered significant and selected for further analysis. Additionally, a random forest algorithm has been used to determine the score of each feature by evaluating its contribution to the accuracy of the random forest model. Once the feature selection process is complete and the most relevant features are identified, SMOTE-NC is applied to generate synthetic samples. This is achieved by considering the local neighborhood of each minority sample and creating new samples that possess similar feature values within that neighborhood. The aim is to augment the minority class and rebalance the dataset. The resulting dataset now consists of the selected relevant features and the synthetic samples from SMOTE-NC. After the feature selection process and the application of SMOTE-NC, these synthetic data, along with the real data, are fed into the discriminator. Now, the discriminator's role is to find the difference between synthetic data and real data. The discriminator takes input from SMOTE-NC and GAN. Training the discriminator with a mixed dataset of real and synthetic samples improves its ability to find out and classify the two types of data accurately. The feedback loop between the discriminator and the generator, responsible for producing synthetic samples, encourages the generator to generate more realistic and representative synthetic samples. Through this adversarial process, the discriminator improves its ability to identify real and synthetic samples. The addition of synthetic samples reduces bias towards the majority class and enhances the model's generalization capabilities, resulting in more dependable and robust outcomes. Finally, the output is obtained after the GAN-generated dataset has passed through a machine-learning classifier. This technique encourages the neural network to learn more robust data representations. This regularization technique prevents the model from overly relying on particular features or patterns within the dataset, thereby improving the model's ability to generate synthetic data.

3.4 Pseudocode for SMOTE_NC-GAN

Our SMOTE_NC-GAN Model, the modified version, incorporates SMOTE-NC to balance the dataset consisting of both continuous and nominal features. The pseudocode for SMOTE_NC-GAN is illustrated in `algorithm 1`. The discriminator receives input from SMOTE-NC and the generator, which helps in balancing the dataset. On the other hand, the generator generates synthetic samples using a noise input z. The GAN algorithm iteratively trains the discriminator and the generator to increase the quality of the generated samples. The updated weights of the discriminator and the generator are stored in Wd and Wg, respectively. Finally, the function returns the accumulated fake examples x0, which are produced by combining the output of the discriminator with the input from SMOTE-NC.

4 Experiment and result analysis

We evaluate the performance of the SMOTE_NC-GAN Model in comparison to SMOTE, non oversampled and SMOTENC on three imbalanced datasets- namely UCI CreditCard Fraud, UCI Adult Income Dataset, and UCI Cervical Cancer Dataset. The datasets were taken from the UCI machine learning repository. Datasets Adult Income, credit card default, and Cervical Cancer are imbalanced with 24.78%, 22.12%, and 6.41% minority class respectively as shown in Table 1. This section details the experimental setup, including dataset description and evaluation metrics.

Algorithm 1 SMOTE_NC-GAN Algorithm.

Initialize parameters:
$T = 100$ ▷ Number of GAN training iterations
$S = 1$ ▷ Step size
$l = 0.001$ ▷ Learning rate
for $t \leftarrow 1$ **to** T **do**
  for $s \leftarrow 1$ **to** S **do**
    $x_{\text{synthetic}} \leftarrow \text{SMOTE-NC}(x_{\text{real}})$
    $D_{\text{real}} \leftarrow \text{discriminator}(x_{\text{real}})$
    $D_{\text{synthetic}} \leftarrow \text{discriminator}(x_{\text{synthetic}})$
    $g_d \leftarrow \text{SGD}(-\log(D_{\text{real}}) - \log(1 - D(G(z))) - \log(1 - D_{\text{synthetic}}), W_d, l)$
    $W_d \leftarrow \text{updateWeights}(g_d, W_d)$
  end for
  $g_g \leftarrow \text{SGD}(-\log(D(G(z))), W_g, \text{l})$
  $W_g \leftarrow \text{updateWeights}(g_g, W_g)$
end for
$y' \leftarrow \text{generated_samples}(\text{Modeld}(W_d, x, \text{SMOTE-NC}(x_{\text{synthetic}}), z), \text{Modelg}(W_g, x, z), n_{\text{fake}})$
Return y'

4.1 Dataset description

For our experiments, we utilized three distinct UCI imbalanced datasets: Credit Card Default dataset, Adult Income dataset, and Cervical Cancer dataset. Each dataset poses unique challenges due to class imbalance. Here is a brief description of each dataset:

1. **U**CI Credit Card Default Dataset: The UCI Credit Card Default dataset consists of 30,000 samples with 24 features. It is commonly used for credit card fraud detection, where the goal is to predict credit card default. The credit card dataset mainly consists of most non-fraudulent transactions, and only a small portion are fraudulent. The dataset provides a challenging scenario with significant class imbalance, making it suitable for evaluating the effectiveness of imbalance handling techniques.
2. **U**CI Adult Income dataset: The UCI Adult Income dataset comprises 48,842 samples with 14 features with 3,620 missing values. It focuses on predicting whether an individual's income exceeds a certain threshold, making it a binary classification problem.
3. **U**CI Cervical Cancer: The UCI Cervical Cancer dataset contains 858 samples with 36 features and is employed in cervical cancer prediction. This dataset poses a significant class imbalance challenge.

Each dataset presents its unique characteristics and challenges in terms of class imbalance, which allows us to measure the performance of SMOTE_NC-GAN Model and SMOTE algorithm in different domains.

Table 1 Dataset Details

Dataset	# of Samples	Minor Samples	Major Samples	Total Features	Minority Class	Description
Credit Card Default	30,000	6,636	23,364	24	22.12%	Identified fraudulent
Adult Income	48,842	11208	34014	14	24.78%	Predicts personal income levels
Cervical Cancer	858	55	803	36	6.41%	Predicts cervical cancer occurrences

4.2 Experimental setup

In this section, the analysis of the experiment of our model, i.e., the SMOTE_NC-GAN model, is done. Our model combines the Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTE-NC) with a Generative Adversarial Network (GAN). We have also introduced a customized GAN loss function to enhance the effectiveness of the combined approach. The experiment optimized parameters for the discriminator, generator, and neural network to achieve the best results, as shown in Table 2.

There are four hidden layers in generator and each hidden layer consists of 256, 128, 64, and 32 neurons, respectively. The discriminator network comprises two layers one is a linear layer, and the other is a leaky-ReLU layer. During GAN training, a binary cross-entropy activation function is used, with a training data batch size of 64 and an initial learning rate of 0.0002. The Adam optimizer is employed for training. After performing the pre-processing steps, SMOTE-NC oversampling is applied with k = 5 neighbors. This oversampling technique helps balance the dataset by generating synthetic samples for minority classes, both for nominal and continuous features. Also, a new loss term is integrated into the discriminator loss called the proposed discriminator loss. This loss enables the GAN to improve the synthetic samples it generates further based on the samples generated by the SMOTE-NC technique. This integration allows the GAN to generate synthetic samples that better reflect the dataset's characteristics, especially for minority classes, and improves the overall performance of the GAN in handling imbalanced data.

4.3 Result analysis

We evaluated the performance of several oversampling techniques, including No Oversampling (Original Data),SMOTE, SMOTE-NC with our proposed method SMOTE_NC-GAN using evaluation metrics.Examining various metrics provides us with a comprehensive understanding of the models' capabilities, encompassing overall accuracy, precision in identifying positive instances, and the balance between precision and recall. Here, we present the experimental results obtained with our proposed method and baseline techniques. The Tables 3, 4, and 5 show the outcomes of the analysis on three imbalanced datasets i.e. the credit card default dataset, adult income dataset and cervical cancer dataset respectively. We employed an NVIDIA 3080 GPU to expedite computational processes. The training times for each dataset were recorded as follows:Adult dataset: 35.2 minutes, Cancer dataset: 11.4 minutes and Credit card fraud dataset: 32.6 minutes.

Table 2 Parameters of SMOTE_NC-GAN Model

Parameters			
Parameter	Discriminator	Generator	Neural Network
Activation	LeakyReLU	LeakyReLU	LeakyReLU
Learning Rate	0.0002	0.0002	0.0002
Layers	5	5	8
Optimizer	Adam	Adam	Adam
Neurons	256,128,64,32	256, 128,64,32	256, 128,128,64,64,32,4,1
Batch Size	64	64	64

Table 3 Comparison Analysis of UCI Credit Card Default on Imbalanced Datasets

Model	Metrics	SMOTE	SMOTE-NC	Non Oversampled	SMOTE_ NC_GAN
Random Forest	**Accuracy**	0.834	0.8373	0.8153	0.9195
	Precision	0.8552	0.8565	0.6391	0.8584
	Recall	0.805	0.811	0.3587	0.814
	F1-Score	0.8294	0.8331	0.4595	0.8356
	AUC	0.9122	0.9137	0.7567	0.9718
Neural Network	**Accuracy**	0.501	0.6711	0.7812	0.7518
	Precision	0.501	0.8203	0	0.6014
	Recall	1	0.4398	0	0.0372
	F1-Score	0.6675	0.5726	0	0.0702
	AUC	0.5	0.758	0.5	0.5964
Logistic Regression	**Accuracy**	0.5756	0.6423	0.7812	0.7605
	Precision	0.6254	0.6438	0	0.6433
	Recall	0.3812	0.6403	0	0.1056
	F1-Score	0.4737	0.642	0	0.1814
	AUC	0.6404	0.6977	0.6566	0.6856

The overall performance of data is measured using precision, recall, F1-score, accuracy, and AUC. From the table, it can be seen that the Proposed SMOTE_NC-GAN outperformed all of the techniques across all measures in all datasets. Based on the data provided, the Random Forest model consistently outperforms both Logistic Regression and Neural Network models across the three datasets: UCI Credit Card, UCI Adult, and UCI Cancer.

Table 4 Comparison Analysis of UCI Adult Income on Imbalanced Datasets

Model	Metrics	SMOTE	SMOTE-NC	Non Oversampled	SMOTE_NC_GAN
Random Forest	**Accuracy**	0.8965	0.89	0.8554	0.9439
	Precision	0.9	0.8841	0.7326	0.8859
	Recall	0.8917	0.897	0.6096	0.8925
	F1-Score	0.8958	0.8905	0.6655	0.8892
	AUC	0.9633	0.9558	0.9029	0.9852
Neural Network	**Accuracy**	0.4989	0.8463	0.764	0.7782
	Precision	0.4989	0.8295	0	0.6462
	Recall	1	0.8709	0	0.2663
	F1-Score	0.6657	0.8497	0	0.3772
	AUC	0.5	0.9244	0.5	0.7766
Logistic Regression	**Accuracy**	0.6845	0.7982	0.8024	0.7446
	Precision	0.6994	0.7668	0.6947	0.465
	Recall	0.6444	0.8559	0.2902	0.0852
	F1-Score	0.6708	0.8089	0.4094	0.144
	AUC	0.7766	0.8592	0.7631	0.7167

Table 5 Comparison Analysis of Cervical Cancer on Imbalanced Datasets

Model	Metrics	SMOTE	SMOTE-NC	Non Oversampled	SMOTE_NC_GAN
Random Forest	**Accuracy**	0.972	0.9783	0.9419	0.9891
	Precision	0.9589	0.9595	0.5714	0.9718
	Recall	0.979	0.993	0.3636	0.9885
	F1-Score	0.9689	0.9759	0.4444	0.9801
	AUC	0.9976	0.9964	0.9647	0.9988
Neural Network	**Accuracy**	0.4441	0.9658	0.936	0.9487
	Precision	0.4441	0.9342	0	0.9434
	Recall	1	0.993	0	0.8621
	F1-Score	0.6151	0.9627	0	0.9009
	AUC	0.5	0.9903	0.8255	0.9883
Logistic Regression	**Accuracy**	0.8665	0.972	0.9535	0.9565
	Precision	0.9032	0.9467	0.6364	0.9398
	Recall	0.7832	0.993	0.6364	0.8966
	F1-Score	0.839	0.9693	0.6364	0.9176
	AUC	0.9478	0.9896	0.8944	0.9874

4.3.1 Dataset-specific observations

Random Forest achieves the highest accuracy (0.9195) and dominates in other metrics like Precision, Recall, F1-score, and ROC-AUC. Logistic Regression and Neural Network have similar accuracies, but Random Forest's performance is significantly superior. Again, Random Forest leads in all metrics, with an accuracy of 0.9439. Neural Network (0.7782) outperforms Logistic Regression (0.7446) in terms of accuracy, but both are substantially behind Random Forest. Random Forest showcases the highest accuracy (0.9891). While Neural Network (0.9487) and Logistic Regression (0.9565) have comparable accuracies, they still lag behind Random Forest. This model has the best performance on the UCI Cancer dataset in terms of accuracy (0.9565), followed by the UCI Credit Card (0.7605) and then the UCI Adult dataset (0.7446). Unsurprisingly, given its robust nature, the Random Forest model performs exceptionally well across all datasets. Its performance on the UCI Cancer dataset stands out with an impressive accuracy of 0.9891. The Neural Network model performs best on the UCI Cancer dataset with an accuracy of 0.9487. It has similar performances on the UCI Credit Card (0.7518) and UCI Adult (0.7782) datasets.

5 Discussion and comparison

The ROC curve of the Adult income Dataset is shown in Figs. 5. Figures 2, 3, and 4 illustrate the confusion matrix between the proposed and existing techniques. These results indicate that our proposed oversampling method, which combines SMOTE-NC with a customized GAN loss, outperforms the baseline oversampling techniques across multiple evaluation metrics for two out of three datasets. SMOTE_NC-GAN as an oversampling technique has resulted in notable enhancements in classification performance. This progress has exhibited promising results, effectively boosting the performance of classification models, particularly in

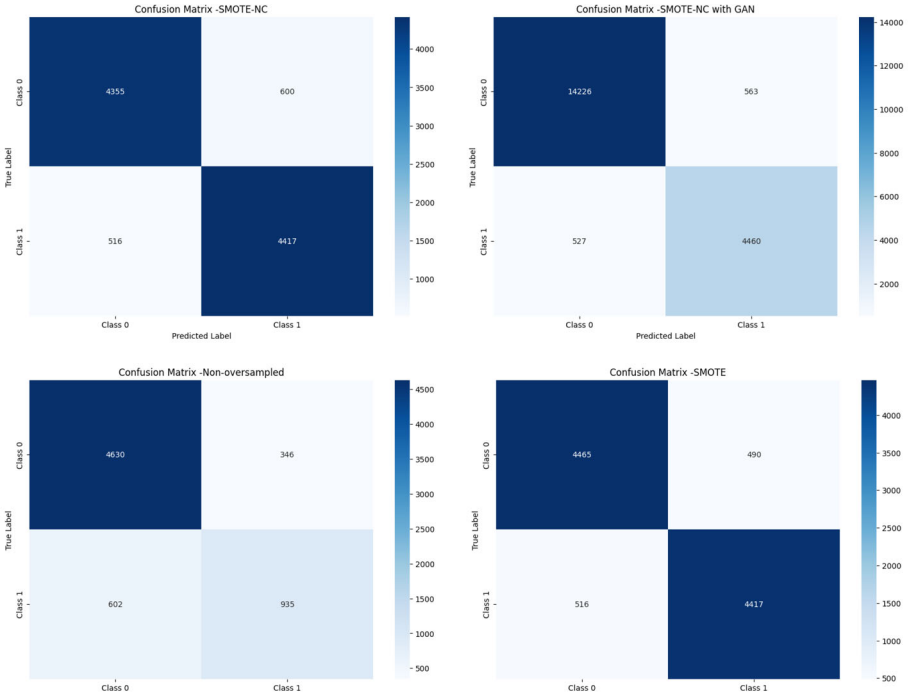


Fig. 2 Performance measurement of without and with oversampling techniques on Adult Dataset using confusion matrices

situations where imbalanced dataset problems are common. For the UCI Credit Card Default dataset, our method achieved an F1 score of 0.8431, which is better compared to the other methods. Similar improvements were observed for the UCI Cervical Cancer datasets. There was no significant improvement observed in the UCI Adult Income dataset; this can result from the fact that the categorical features in this dataset are less correlated, and that's why SMOTE outperforms every other method. The customized GAN loss function contributed to generating both balanced and realistic synthetic samples, resulting in improved model performance.The performance of model is measured using precision, recall, F1-score, accuracy, and AUC. When comparing with the SMOTified-GAN model proposed by Sharma [65], it was found that the F1-score of the SMOTE_NC-GAN model on the credit card dataset was 91.95, which outperformed the F1-score of 82.43 achieved by SMOTified-GAN on the same dataset. This indicates that the proposed model performed better on the credit card dataset. Similarly, when comparing with SMOTE-Voting-PCA [66] on the Cervical Cancer dataset, it was observed that the SMOTE_NC-GAN model achieved higher accuracy and F1 score (98.91 and 98.01 respectively) compared to the accuracy and F1 score (92.89 and 93.35). This suggests the superior performance of the SMOTE_NC-GAN model on the Cervical Cancer dataset. In the Adult Income dataset, the accuracy of the proposed SMOTE_NC-GAN model was 94.39, which surpassed the accuracy of 87.3 in [67]. This comparison underscores the effectiveness of incorporating both SMOTE-NC and GAN techniques in enhancing model performance across different datasets. SMOTE_NC-GAN model consistently outperforms various baseline techniques across UCI Credit Card Default, Adult Income, and Cervical Cancer datasets, particularly with Random Forest. The SMOTE-NC-GAN model achieves

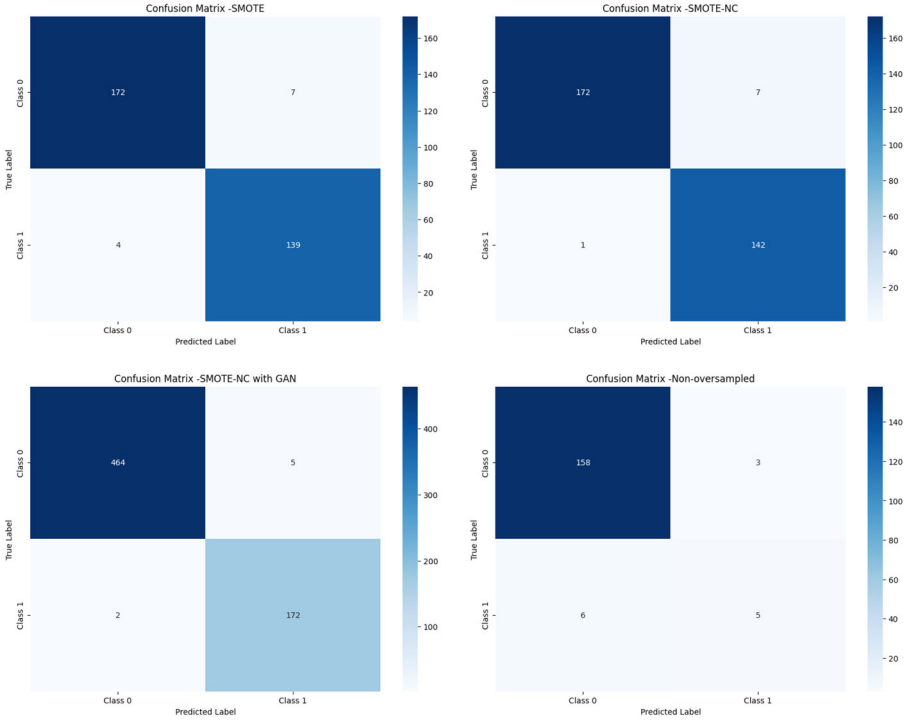


Fig. 3 Performance measurement of without and with oversampling techniques on Cancer Dataset using confusion matrices

an F1-score of 91.95 on the credit card dataset, underscoring its efficacy in addressing imbalanced data. In the UCI Adult Income dataset, maintaining a strong accuracy of 94.39 underscores the model's robustness. In summary, the evaluation metrics confirm the effectiveness of the SMOTE_NC-GAN Model in handling imbalanced datasets. The results offer valuable insights into the model's performance, further supporting its potential as a robust solution for addressing imbalanced classification tasks.

6 Limitations

Despite the promising outcomes achieved by employing the SMOTE_NC-GAN Model to tackle data imbalance in text classification, there exist several limitations that need to be addressed in future research:

- SMOTE_NC-GAN introduces additional computational overhead. The process of generating synthetic samples using GANs and conducting oversampling with SMOTE-NC can notably amplify the computational resources needed, especially for large-scale text datasets.
- The performance of the proposed model is highly dependent on the selection of hyperparameters, which encompass parameters related to SMOTE-NC, GAN architecture, and text classification algorithms. Suboptimal hyperparameter configurations may result in reduced classification accuracy or training instability.

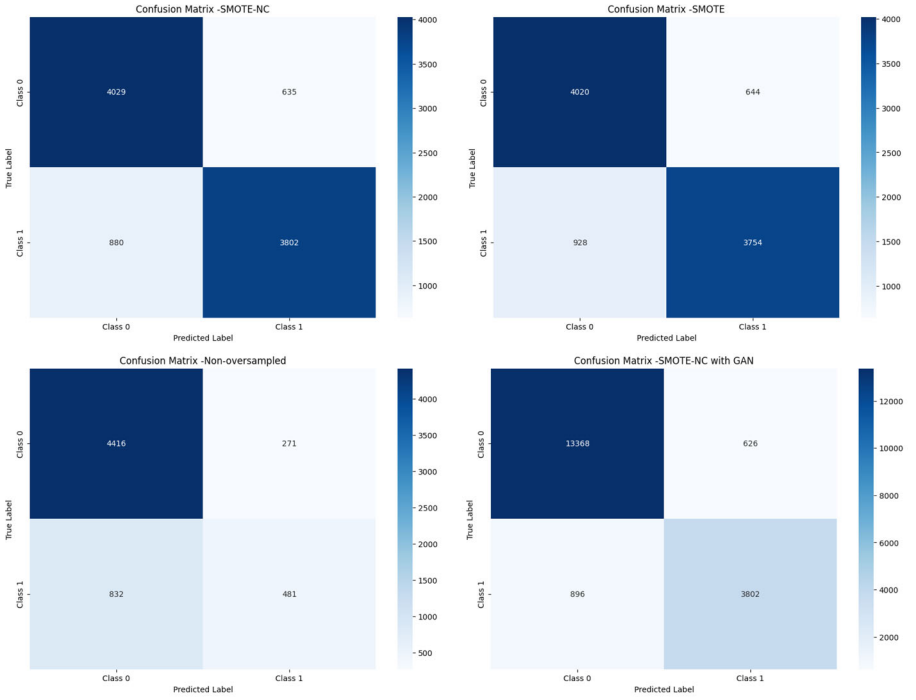


Fig. 4 Performance measurement of without and with oversampling techniques on Credit Card using confusion matrices

- SMOTE_NC-GAN model is effective in addressing class imbalance within the training data, its capacity to generalize to unseen or out-of-distribution samples is restricted. Investigating the model's adaptability to various text domains or languages, especially those with substantially distinct characteristics from the training data, is an avenue for future exploration.

7 Conclusion and future work

Our research introduces and evaluates the performance of SMOTE_NC-GAN Model approaches for handling class imbalanced problems using three imbalanced datasets in mixed data (continuous data, categorical data). Our method demonstrates significant improvements in F1 score, Precision, and recall compared to conventional oversampling techniques for the UCI Credit Card Default, UCI Adult Income, and UCI Cervical Cancer datasets. The results demonstrate that our proposed method outperforms SMOTE and SMOTE_NC across various statistical measures.The findings of our proposed techniques provide valuable insights to researchers and practitioners, emphasizing the advantages of GAN-based approaches and their potential to address marginalization and sensitivity issues associated with existing oversampling techniques. In the future, we aim to extend our research by employing the SMOTE_NC-GAN Model to tackle class imbalanced challenges in diverse,

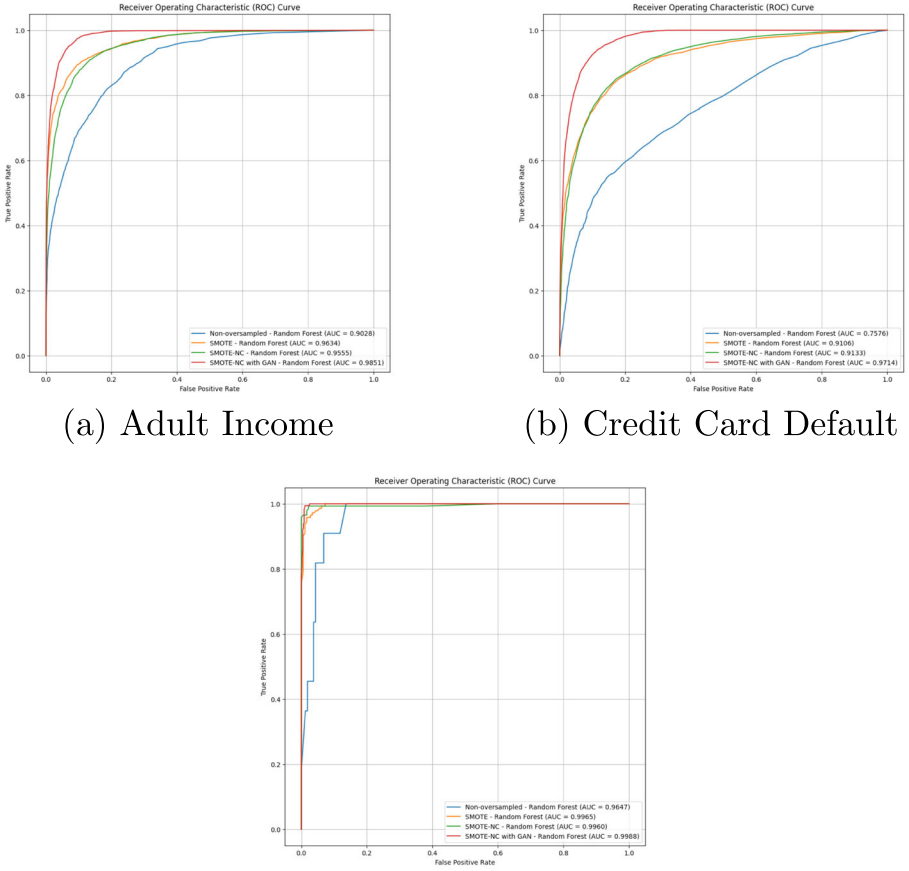


Fig. 5 True Positive rate vs False Positive rate for AUC Curves

highly imbalanced, and large-scale datasets. Moreover, we will conduct thorough experimentation and fine-tune the model parameters to optimize its effectiveness and overall performance.

Acknowledgements Acknowledgments We would like to thank the anonymous reviewers for their valuable comments, which helped us to improve the organization, content, and quality of this Manuscript

Data availability The dataset for this research work are available at
- https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients
- https://archive.ics.uci.edu/dataset/2/adult
- https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors

Declarations

Conflicts of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Jiang M, Liang Y, Feng X, Fan X, Pei Z, Xue Y, Guan R (2018) Text classification based on deep belief network and softmax regression. Neural Comput Appl 29:61–70
2. Japkowicz N, Stephen S (2002) The class imbalance problem: A systematic study. Intell Data Anal 6(5):429–449
3. Fong S, Wong R, Vasilakos AV (2015) Accelerated pso swarm search feature selection for data stream mining big data. IEEE Trans Serv Comput 9(1):33–45
4. Onan A, Korukoğlu S, Bulut H (2016) Ensemble of keyword extraction methods and classifiers in text classification. Expert Syst Appl 57:232–247
5. Diab DM, El Hindi KM (2017) Using differential evolution for fine tuning naïve bayesian classifiers and its application for text classification. Appl Soft Comput 54:183–199
6. Makki S, Assaghir Z, Taher Y, Haque R, Hacid M-S, Zeineddine H (2019) An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access 7:93010–93022
7. Khairy M, Mahmoud TM, Abd-El-Hafeez T (2024) The effect of rebalancing techniques on the classification performance in cyberbullying datasets. Neural Comput Appl 36(3):1049–1065
8. Khairy M, Mahmoud TM, Omar A, Abd El-Hafeez T (2023) Comparative performance of ensemble machine learning for arabic cyberbullying and offensive language detection. Lang Resour Eval 1–18
9. Bria A, Marrocco C, Tortorella F (2020) Addressing class imbalance in deep learning for small lesion detection on medical images. Comput Biol Med 120:103735
10. Qin Z, Liu Z, Zhu P, Xue Y (2020) A gan-based image synthesis method for skin lesion classification. Comput Methods Programs Biomed 195:105568
11. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G (2017) Learning from class-imbalanced data: Review of methods and applications. Expert Syst Appl 73:220–239
12. Zhao L, Zhang L, Wu Z, Chen Y, Dai H, Yu X, Liu Z, Zhang T, Hu X, Jiang X et al (2023) When brain-inspired ai meets agi. Meta-Radiology 100005
13. Yang S, Linares-Barranco B, Chen B (2022) Heterogeneous ensemble-based spike-driven few-shot online learning. Front Neurosci 16:850932
14. Yang S, Tan J, Chen B (2022) Robust spike-based continual meta-learning improved by restricted minimum error entropy criterion. Entropy 24(4):455
15. Yang S, Chen B (2023) Snib: improving spike-based machine learning using nonlinear information bottleneck. IEEE Transactions on Systems, Man, and Cybernetics: Systems
16. Yang S, Wang H, Chen B (2023) Sibols: robust and energy-efficient learning for spike-based machine intelligence in information bottleneck framework. IEEE Trans Cogn Dev Syst
17. Yang S, Chen B (2023) Effective surrogate gradient learning with high-order information bottleneck for spike-based machine intelligence. IEEE Trans Neural Netw Learn Syst
18. Ling CX, Sheng VS (2008) Cost-sensitive learning and the class imbalance problem. Encycl Mach Learn 2011:231–235
19. Mariani G, Scheidegger F, Istrate R, Bekas C, Malossi C (2018) Bagan: Data augmentation with balancing gan. arXiv preprint arXiv:1803.09655
20. Hoens TR, Qian Q, Chawla NV, Zhou Z-H (2012) Building decision trees for the multi-class imbalance problem. In: Advances in Knowledge Discovery and Data Mining: 16th Pacific-Asia Conference, PAKDD 2012, Kuala Lumpur, Malaysia, May 29-June 1, 2012, Proceedings, Part I 16, pp 122–134. Springer
21. Guo X, Yin Y, Dong C, Yang G, Zhou G (2008) On the class imbalance problem. In: 2008 Fourth International Conference on Natural Computation, vol 4, pp 192–201. IEEE
22. Japkowicz N, Stephen S (2002) The class imbalance problem: A systematic study. Intell Data Anal 6(5):429–449
23. Johnson JM, Khoshgoftaar TM (2019) Survey on deep learning with class imbalance. J Big Data 6(1):1–54
24. Kotsiantis S, Kanellopoulos D, Pintelas P et al (2006) Handling imbalanced datasets: A review. GESTS Int Trans Comput Sci Eng 30(1):25–36
25. Fernández A, Garcia S, Herrera F, Chawla NV (2018) Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. J Artif Intell Res 61:863–905
26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
27. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. Commun ACM 63(11):139–144
28. He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284
29. Prati RC, Batista GE, Silva DF (2015) Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowl Inf Syst 45:247–270

30. Cheng F, Zhang J, Wen C (2016) Cost-sensitive large margin distribution machine for classification of imbalanced data. Pattern Recogn Lett 80:107–112
31. López V, Fernández A, Del Jesus MJ, Herrera F (2013) A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets. Knowl-Based Syst 38:85–104
32. Polikar R (2012) Ensemble learning. Ensemble machine learning: Methods and applications 1–34
33. García V, Sánchez JS, Mollineda RA (2012) On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. Knowl-Based Syst 25(1):13–21
34. Yap BW, Rani KA, Rahman HAA, Fong S, Khairudin Z, Abdullah NN (2014) An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In: Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), pp 13–22. Springer
35. Kubat M, Matwin S et al (1997) Addressing the curse of imbalanced training sets: one-sided selection. In: Icml, vol 97, p 179. Citeseer
36. Tomek I (1976) Two modifications of cnn
37. Gowda K, Krishna G (1979) The condensed nearest neighbor rule using the concept of mutual nearest neighborhood (corresp.). IEEE Trans Inf Theory 25(4):488–490
38. Niemann H, Goppert R (1988) An efficient branch-and-bound nearest neighbour classifier. Pattern Recogn Lett 7(2):67–72
39. Mamdouh Farghaly H, Abd El-Hafeez T (2023) A high-quality feature selection method based on frequent and correlated items for text classification. Soft Comput 27(16):11259–11274
40. Mamdouh Farghaly H, Abd El-Hafeez T (2022) A new feature selection method based on frequent and associated itemsets for text classification. Concurrency and Computation: Practice and Experience 34(25):7258
41. Hassan E, Abd El-Hafeez T, Shams MY (2024) Optimizing classification of diseases through language model analysis of symptoms. Sci Rep 14(1):1507
42. Mahmoud T, Abd-El-Hafeez T, El-Deen D (2016) A design of an automatic web page classification system. British J Appl Sci Technol 18(6):1–14
43. Khairy M, Mahmoud TM, Abd-El-Hafeez T (2021) Automatic detection of cyberbullying and abusive language in arabic content on social networks: a survey. Procedia Comput Sci 189:156–166
44. Koshiry AME, Eliwa EHI, Abd El-Hafeez T, Omar A (2023) Arabic toxic tweet classification: Leveraging the arabert model. Big Data Cogn Comput 7(4):170
45. Omar A, Abd El-Hafeez T (2023) Quantum computing and machine learning for arabic language sentiment classification in social media. Sci Rep 13(1):17305
46. Omar A, Mahmoud TM, Abd-El-Hafeez T (2018) Building online social network dataset for arabic text classification. In: The international conference on advanced machine learning technologies and applications (AMLTA2018), pp 486–495. Springer
47. Omar A, Mahmoud TM, Abd-El-Hafeez T, Mahfouz A (2021) Multi-label arabic text classification in online social networks. Inf Syst 100:101785
48. Chawla NV, Japkowicz N, Kotcz A (2004) Special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsl 6(1):1–6
49. Douzas G, Bacao F, Last F (2018) Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. Inf Sci 465:1–20
50. Kovács G (2019) Smote-variants: A python implementation of 85 minority oversampling techniques. Neurocomputing 366:352–354
51. Dattagupta SJ (2018) A performance comparison of oversampling methods for data generation in imbalanced learning tasks. PhD thesis
52. Cheng K, Zhang C, Yu H, Yang X, Zou H, Gao S (2019) Grouped smote with noise filtering mechanism for classifying imbalanced data. IEEE Access 7:170668–170681
53. Lusa L et al (2012) Evaluation of smote for high-dimensional class-imbalanced microarray data. In: 2012 11th International Conference on Machine Learning and Applications, vol 2, pp 89–94. IEEE
54. Douzas G, Bacao F (2018) Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Syst Appl 91:464–471
55. Wang X, Man Z, You M, Shen C (2017) Adversarial generation of training examples: Applications to moving vehicle license plate recognition. arXiv:1707.03124
56. Bin H, Weihai C, Xingming W, Chun-Liang L (2017) High-quality face image sr using conditional generative adversarial networks. arXiv preprint arXiv:1707.00737
57. Antipov G, Baccouche M, Dugelay J-L (2017) Face aging with conditional generative adversarial networks. In: 2017 IEEE International Conference on Image Processing (ICIP), pp 2089–2093. IEEE

58. Mariani G, Scheidegger F, Istrate R, Bekas C, Malossi C (2018) Bagan: Data augmentation with balancing gan. arXiv preprint arXiv:1803.09655
59. Berthelot D, Schumm T, Metz L (2017) Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717
60. Odena A, Olah C, Shlens J (2017) Conditional image synthesis with auxiliary classifier gans. In: International conference on machine learning, pp 2642–2651. PMLR
61. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. Adv Neural Inf Process Syst 29
62. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: International Conference on Machine Learning, pp 214–223. PMLR
63. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196
64. Salehinejad H, Valaee S, Dowdell T, Colak E, Barfett J (2018) Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 990–994. IEEE
65. Sharma A, Singh PK, Chandra R (2022) Smotified-gan for class imbalanced pattern classification problems. Ieee Access 10:30655–30665
66. Alsmariy R, Healy G, Abdelhafez H (2020) Predicting cervical cancer using machine learning methods. Int J Adv Comput Sci Appl 11(7)
67. Alex SA, Nayahi JJV (2023) Classification of imbalanced data using smote and autoencoder based deep convolutional neural network. Internat J Uncertain Fuzziness Knowledge-Based Systems 31(03):437–469

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT) | 979-8-3503-7131-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/InCACCT61598.2024.10551131

Multimodal Architecture for Sentiment Recognition via employing Multiple Modalities

Samridhi Kapoor
Department of Information Technology
Delhi Technological University
Bawana Road, Delhi, India
samridhi.kapoor2903@gmail.com

Shubham Gulati
Department of Information Technology
Delhi Technological University
Bawana Road, Delhi, India
shubham53.sg@gmail.com

Sumit Verma
Department of Information Technology
Delhi Technological University
Bawana Road, Delhi, India
vermasumit923@gmail.com

Ananya Pandey
Department of Information Technology
Delhi Technological University
Bawana Road, Delhi, India
ananyaphdit08@gmail.com

Dinesh Kumar Vishwakarma
Department of Information Technology
Delhi Technological University
Bawana Road, Delhi, India
dvishwakarma@gmail.com

***Abstract*—In the past, a lot of research has been done on text-driven sentiment analysis using benchmark multimodal Twitter-15 and Twitter-17 combined dataset. A small number of relevant studies have also shown the use of visual analysis to forecast sentiment in pictures. However, the majority of the research has only examined one modality of data—text, photos, or GIF videos. Lately, with photos, memes, and GIFs taking over social media feeds, typographic and info-graphic visual material has emerged as a significant component of social media. The suggestion is a multi-modal sentiment analysis model that may be used to ascertain the sentiment polarity and score of each incoming tweet, taking into the consideration of text as well as image features of the tweet. Text-based sentiment scoring is done using BERT, RoBERTa, XLNet. Vision-based sentiment scoring is done using ResNet-50, RegNet, ResNeXt. The study came up with a Multi-modal Sentiment Recognition model consisting of text and image models and combining the sentiment scores from the separately processed text and image. Using the Twitter-1517 benchmark multi-modal twitter dataset, the model with RegNet for visual scoring and RoBERTa for text scoring demonstrates a high performance accuracy of 77.24%. The study goes on to show that integrating text and picture characteristics performs better than independent models that only use text analysis or images.**

Keywords— Sentiment Recognition (SR), Multi-modal Sentiment Recognition (MSR), RoBERTa, RegNet

I. INTRODUCTION

Today's world with the ever-expanding digitalisation, social media platforms have become veritable treasure troves of human expression, where sentiments are shared openly and rapidly. As fig 1 displays, a significant quantity of opinionated social big data is available on the present, widely available, and reasonably priced Web generation, which helps with decision-making. Sentiment analysis, which tracks people's views and moods by evaluating their informal, unstructured, high-dimensional, and noisy multimodal social data, has become more important in a short time. The formal definition of sentiment analysis is the computer examination of people's feelings, attitudes, and beliefs concerning an object. Conventional sentiment analysis mostly uses one of the following modalities: voice, picture, or text. Humans may, however, artificially communicate their feelings in a variety of formats, including words, voice, body language, and facial expressions. As such, relying just on one modality for sentiment analysis has distinct drawbacks. One of the better ways to get over this restriction is to use multimodal data synthesis to assess attitudes, which is motivating more and more academics to work toward it. To specifically address the aforementioned drawbacks in single-modality sentiment analysis, a number of techniques are investigated, including modality representation, modality alignment, and modality fusion.

Fig. 1 Samples with the predicted class labels on the combined dataset of Twitter-1517

This kind of NLP (Natural Language Processing) challenge is based on analysing the vast volume of user-generated online information that is created on a regular basis in user-editable web forums, blogs, social networks, and e-commerce sites. Relevant research in the field describes how several methods—like lexicon-based, hybrid, machine learning, and concept-based (ontology and contextual and -based) approaches—are used to generate opinions.

A. Sentiment Analysis: A Complex Terrain

Leveraging computational technology, sentiment analysis (SA) involves the examination of data, including images, audio, or textual data, to deduce the conveyed emotions or states of emotion. Some other terms for SA are Emotion mining, Mood tracking, etc.

The aim of this noteworthy research endeavour in natural language processing is to enable computers to understand and grasp human emotions.

979-8-3503-7131-4/24/$31.00 ©2024 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 24,2024 at 04:33:32 UTC from IEEE Xplore. Restrictions apply.

Sentiment analysis includes sentiment mining, opinion extraction, sentiment categorization, and emotion analysis, among other things. Of these, sentiment categorization is the topic that has been explored the most. Based on different degrees of granularity, sentiment categorization may be further divided into sentence level, document level, and aspect level categories as mentioned in [1]. The objective of document level sentiment recognition is to showcase the overall polarity of sentiment of a particular document or even a long segment of text. Sentiment classification at the sentence level predicts the sentiment polarity of each sentence. Sentiment classification at the aspect level is distinct from that at the document or sentence levels since it focuses more on predicting the inclination of sentiment of certain target features. It examines feelings pertaining to particular objectives, motifs, or parts of the text.

B. Multi-modal Sentiment Recognition

Sentiment recognition using various modalities (text, picture, etc.) is called multimodal sentiment recognition (MSR)[2], [3], [4], [5], [6], [7]. Multimodal sentiment analysis is able to gather more accurate and complete sentiment information from several perception channels as compared to single mode sentiment analysis. For instance, if the line "The weather today is truly amazing!" is examined only from the text, it conveys a good feeling; yet, if the sentence is coupled with a picture of a rainy evening, the entire sentiment becomes sardonic and gloomy. It is challenging to infer the sentiment in this case just from the textual mode. Of course, multimodality can occasionally produce some noise. When connecting, for instance, a text with the sentiment "gloomy" or "depressed" with a picture of a laughing child that conveys a positive attitude, the visual modalities might introduce some noise, which can cause the sentiment to be predicted incorrectly overall.

II. LITERATURE REVIEW

A. Unimodal Sentiment Analysis

Social media is abundant in visual and textual information presented together or in isolation. Sentiment Anaylsis done on the basis of text (as a modality) was known as Text-based Sentiment Recognition (TBSR). Such kind of Analysis can be seen in [8] proposed by Nandwani et al. The paper [12] showcased sentiment (emotion) recognition from text via various emotion models. Consequently, photos could also be fed into a model that analyzed emotions. This type of Sentiment Recognition was referred as Visual Sentiment Recognition. This analysis was done in [9] where Song et al. presented a novel architecture and created a combination of visual attention and CNN.

B. Multi-modal Sentiment Analysis

Multi-modal Sentiment Analysis comprised of analysing sentiments over a dataset involving more than one modalities. It could be a combination of text, voice and image or combination of GIFs and text or any other combination. Ju et al, one of the authors of the paper [10] proposed an innovative learning approach Term Extraction and Sentiment Recognition through Aspect Analysis. The authors in the paper[11] presented a novel application of BERT to target-oriented multimodal sentiment categorization. It had shown that concentrating on certain goals, the authors' contributions tackled the challenging problem of sentiment analysis and improved BERT's ability to handle multimodal data. An innovative architecture was presented in paper [2] that combined Convolution Block Attention Module (CBAM with ConvNet) and BERT model to construct a visual-attention based module. In the paper [12], a study on predicting the emotion tag was done on Tumblr posts which were treated as "self reported emotions". The sentiment possibilities were only two, i.e. positive and negative.

III. PROPOSED MODEL

The following segment describes how textual and visual modalities is used to perform sentiment analysis as well as contrast them with only text and only image methods of performing sentiment analysis. **Error! Reference source not found.** shows the proposed model architecture.

A. Problem Statement

Given a set of multi-modal dataset comprising a tuple of text and image modality for each sample. There are three possibilities for the labels: {positive, neutral, negative}. The goal of MSR is to forecast the emotion connected to a certain sample that hasn't been noticed before.

B. Text-based Sentiment Recognition

Implemented three models for Text-based Sentiment Recognition (TBSR), namely, BERT, RoBERTa and XLNet for text only modality.

Bidirectional Encoder Representations or BERT [13], from Transformers, is a pre-trained natural language processing model that has greatly improved the state-of-the-art in a number of language understanding tasks. In 2018, it was presented by Google researchers.

BERT is based on the Transformer architecture, given by Vaswani et al. in the paper [14]. This architecture utilizes a self-attention mechanism to enclose contextual relationships between words in a sequence. BERT, being a bidirectional model, differs from traditional language models by considering both left and right context during training.

The RoBERTa [15] model and the BERT model have the same architecture. The RoBERTa is just a reimplementation of BERT model with a small changes to hyperparameters and few small adjustments in embedding.General fine-tuning and pre-training operations were carried out on the BERT. On performing these operations in BERT with the exception of the output layers, employs the identical designs. For various downstream tasks, models were prepared utilizing the same pre-trained model parameters. Fine-tuning involved adjusting each and every parameter.

XLNet [16] is an autoregressive Transformer that aims to circumvent the drawbacks of auto encoding and autoregressive language modeling by combining the best of both approaches. Considering all the possibilities of permutations in the factorization order, the XLNet achieves a greater log probability for a given sequence. While considering permutations, all the tokens (taken from left as well as right) is to be included in context for each position. Every position gains the ability to use contextual data from every position, enclosing context in both directions.

Inspired by the recent development of a language model based on pre-trained transformers, processed textual data from Twitter-1517 (combination of Twitter-15 and Twitter-17) dataset using the RoBERTa model. The text-based data is first cleaned and preprocessed to remove unnecessary words and stop words from the sentences. The obtained cleaned text is converted into a collection of tokens—which are just sets of numeric values—by the pre-trained RoBERTa tokenizer. For feature extraction, tokenized text is passed to the feature extractor of RoBERTa, then a layer of Conv1D. The activation function LeakyRelu is used by the text model. Out of the three models, it was evident that RoBERTa performed best.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 24,2024 at 04:33:32 UTC from IEEE Xplore. Restrictions apply.

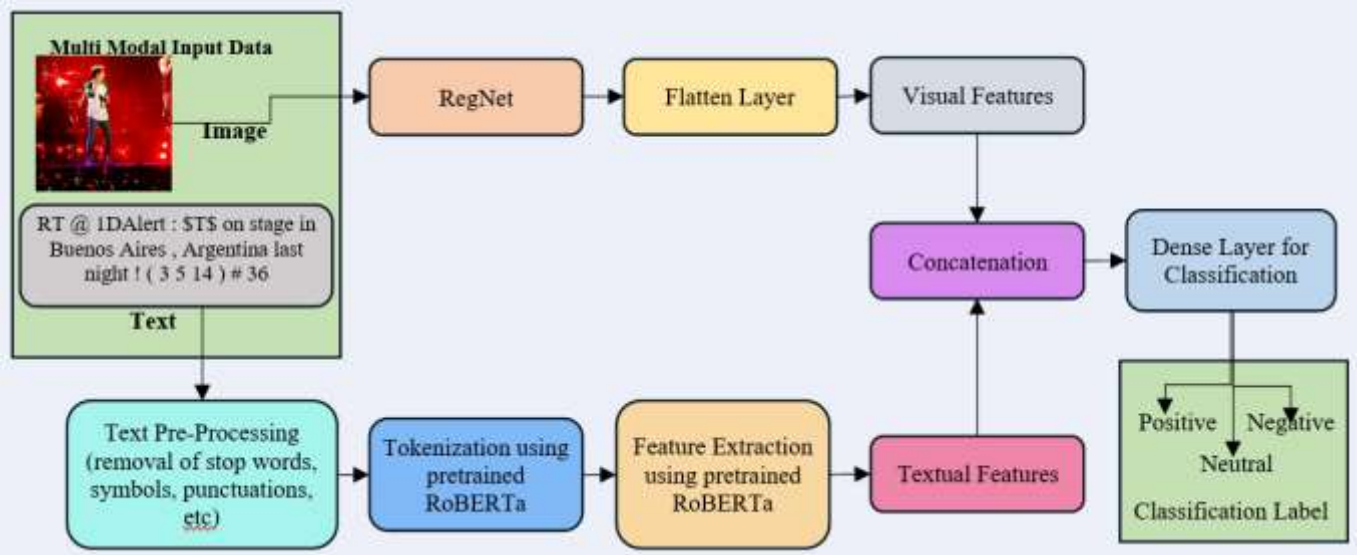


Fig. 2 Proposed Model Architecture

C. *Visual-based Sentiment Recognition*

This section is aimed at providing an overview of the models chosen for vision based sentiment recognition namelyResNet-50[17], RegNet[18] and ResNeXt[19] for Visual-based Sentiment Recognition.

In 2015, Microsoft Research led the way in developing ResNet50, an innovative deep convolutional neural network architecture. This 50-layer iteration of the established ResNet design empowers the assimilation of significantly more complex structures, effectively addressing the challenge posed by vanishing gradients. The ResNet50 framework comprises four integral components: identity block, convolutional block, fully connected layers, and convolutional layers. The identity and convolutional blocks manipulate and refine features extracted by the convolutional layers from the input image, while the fully connected layers execute the ultimate classification. ResNet50 showcases an error rate on par with human performance, establishing it as a formidable model for diverse image classification tasks, encompassing object detection, facial recognition, and medical image analysis. Trained on the extensive ImageNet dataset, ResNet50 has also been adeptly employed as a feature extractor in various applications, including semantic segmentation and object detection.

RegNet introduces a breakthrough in enhancing ResNet and its variants by integrating a regulator module, addressing limitations in conventional shortcut connections. The regulator module employs convolutional Recurrent Neural Networks (RNNs) to extract spatio-temporal information and explore complementary features, overcoming the additive nature of traditional shortcuts. Experimental results on diverse image classification datasets demonstrate RegNet's promising performance, outperforming standard ResNet and other state-of-the-art architectures. This signifies RegNet's potential to advance computer vision tasks by actively extracting and leveraging nuanced features for improved model accuracy and generalization.

The ResNeXt architecture is very similar to that of the ResNet, the sole distinction between the ResNet and ResNeXt architectures is the addition of a new dimension known as the Cardinality dimension. The primary goal of ResNeXt is to manage massive inputs and enhance the network's accuracy without the need to construct deeper layers by raising cardinality, which keeps the complexity comparatively lower.

Inspired by the recent development of a language model based on pre-trained transformers, processed visual-based data for the dataset using the RegNet model. Out of the three, it was evident that RegNet performed best.

D. *Multi-modal Sentiment Analysis*

Textual attributes derived from text-centric model, embodied by RoBERTa, are integrated with output characteristics stemming from the image-oriented model generated by RegNet. The Concatenate layer is used to concatenate the outputs from the text and image models along axis 1. The model may utilize the features that were derived from both modalities thanks to this fusion stage. Following the concatenation, the combined features pass through a dense layer with 512 units and rectified linear unit (ReLU) activation. This layer facilitates the integration of high-level features extracted from both textual and image data. The final layer consists of a dense layer with three units and softmax activation, producing the model's output. The softmax activation ensures that the model provides probability distributions across the three possible classes. Using the last dense layer on this combined feature vector, the emotion polarity is then classified as positive, negative, or neutral.

IV. Experimental Results

A. *Dataset*

Twitter-15 and Twitter-17 as mentioned in [11] are used combine them to form Twitter-1517 to evaluate the approach for the MSR problem. Twitter-1517 is made up of tweets with a photo and a caption. Finding out if the sentiment is positive, neutral or negative is the aim. In the Twitter-1517 dataset, 20 percent of the 11319 total samples are selected for testing.

B. *Implementation Details*

The framework for tokenization and feature extraction uses a RoBERTa-based model for text data and the RegNet module for image data. After optimizing the downstream tasks on the training data, lock all the hyperparameters and modify them for 50 epochs. At this point, the batch size is set to 32 and the Adam optimizer's default learning rate is set to 0.001. Both Adam and the categorical cross-entropy loss function were used as optimizers. TensorFlow and Keras were the frameworks used in the implementation of each model. A Persistence-M GPU is used for the experiments.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 24,2024 at 04:33:32 UTC from IEEE Xplore. Restrictions apply.

Table I. Statistical information for all the dataset used to validate proposed model

Dataset	Number of samples			
	Neutral	Positive	Negative	Total
Twitter-15	3169	1548	630	5347
Twitter-17	2728	2516	728	5972
Twitter-1517	5897	4064	1358	**11,319**

C. *Evaluation Metrics*

To address the given MSR problem statement, have leveraged the Macro-Recall, Accuracy, Macro-precision and Macro-F1 for the assessment of performance of this proposed model.

D. *Main Results*

Compared the results of Twitter-15 and Twitter-17 with the most recent cutting-edge baselines for image-text pairs to demonstrate the strength of proposed approach in Table III and Table IV. Graphical representation of the same is represented in Fig. 3 and **Error! Reference source not found.**. Furthermore, to thoroughly evaluate the effectiveness of model, conducted testing on the combined dataset. The results of this testing can be found in Table V. **Error! Reference source not found.** depicts the loss and accuracy curves for the Twitter-1517 dataset shows the efficacy of proposed method.

Table II. Comparing the accuracy of different models on the Twitter-1517 dataset, RegNet excelled with image data while RoBERTa performed strongly with text data.

Modalities	Twitter-1517		
	Methods utilized		
***Image* (only)**	**ResNet-50**	**ResNeXt**	**RegNet**
	73.84	74.02	**75.81**
	Methods utilized		
Text (only)	**BERT**	**XLNet**	**RoBERTa**
	72.08	70.23	**75.91**

Table III. Comparison of proposed model with the baseline method for Twitter-15 dataset. The comparison graph of the same is shown in Fig. 3.

Baselines	Twitter-15			
	Accuracy	Macro-precision	Macro-Recall	Macro-F1
ESAFN [20]	73.38	-	-	67.37
TomBERT [11]	77.15	-	-	71.75
HIMT [21]	78.14	-	-	73.68
EF-Net [22]	73.65	-	-	-
ModalNet [23]	79.03	-	-	72.50
VAuLT [24]	75.60	-	-	70.0
CapBERT[25]	77.92	-	-	73.90
Proposed (Ours)	**79.24**	**76.13**	**75.94**	**76.03**

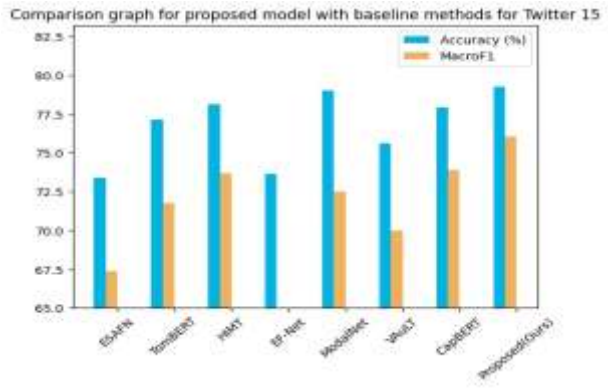


Fig. 3 Comparison Graph of proposed model with baseline methods for Twitter 15

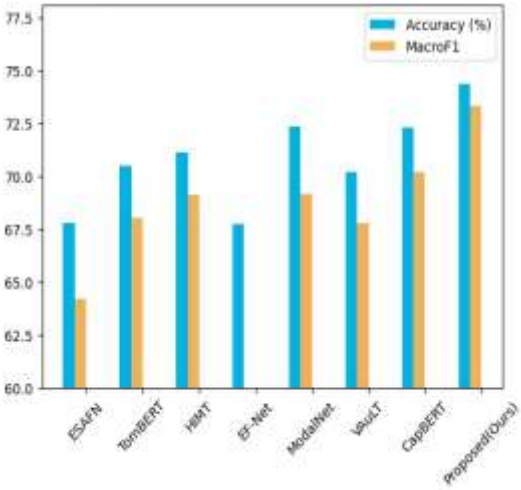


Fig. 4 Comparison Graph of proposed model with baseline methods for Twitter 17

Table IV. Comparison of proposed model with the baseline method for Twitter-17 dataset. The comparison graph of the same is shown in Error! Reference source not found.

Baselines	Twitter-17			
	Accuracy	Macro-precision	Macro-Recall	Macro-F1
ESAFN [20]	67.83	-	-	64.22
TomBERT [11]	70.50	-	-	68.04
HIMT [21]	71.14	-	-	69.16
EF-Net [22]	67.77	-	-	-
ModalNet [23]	72.36	-	-	69.19
VAuLT [24]	70.20	-	-	67.80
CapBERT[25]	72.30	-	-	70.20
Proposed (Ours)	**74.39**	**73.01**	**73.71**	**73.35**

Table V. Final results on the combined dataset for multimodal samples

Modalities	Twitter-1517			
	Methods utilized RoBERTa + RegNet			
Text + Images	**Accuracy**	**Macro-precision**	**Macro-Recall**	**Macro-F1**
	77.24	**77.74**	**77.38**	**76.17**

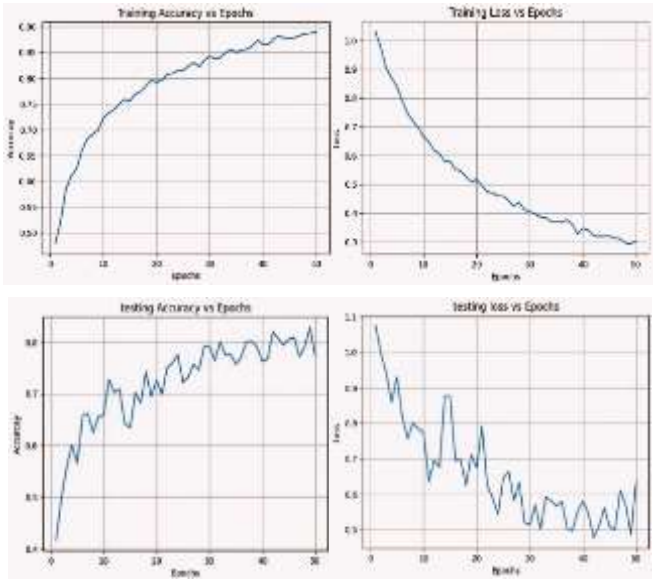


Fig. 5 The loss and accuracy curves for the Twitter-1517 dataset shows the efficacy of proposed method

V. Conclusion And Future Scope

Advancements in multimodal sentiment analysis have the potential to bring about transformative effects across various

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 24,2024 at 04:33:32 UTC from IEEE Xplore. Restrictions apply.

fields, leveraging the combined analysis of textual and visual information. The integration of both modalities allows for a more nuanced understanding of human sentiment expression, enabling applications in diverse domains. Here are several fields that stand to benefit from these advancements: Social Media and Online Platforms, Market Research and Business Intelligence, Healthcare and Well-being, Education and E-Learning, Human-Computer Interaction (HCI), Entertainment and Media, Law Enforcement and Security, Customer Service and Support, Smart Cities and Urban Planning and Cultural and Societal Analysis.

As technology continues to advance and multimodal sentiment analysis techniques mature, the potential applications across these diverse fields are likely to expand, fostering innovation and improvements in decision-making processes.

Unexplored frontiers in multimodal sentiment analysis include the intricate dynamics of cross-modal interactions and dependencies, particularly investigating how changes in one modality influence sentiment expression in another. Cultural and contextual variations in sentiment expression across diverse modalities demand deeper exploration to ensure models can adapt to nuanced expressions. The challenge of real-time analysis in streaming multimodal data, especially in applications like live events or news monitoring, remains an underexplored domain. Ethical considerations, encompassing bias, fairness, and privacy in multimodal sentiment analysis, require further investigation to develop responsible and transparent systems. Integrating human-centric design principles, ensuring model interpretability, and exploring temporal dynamics of emotions across modalities are crucial for user trust and understanding. Extending multimodal sentiment analysis to languages with limited resources and less commonly analysed modalities, such as gestures or haptic feedback, is an ongoing challenge. Furthermore, research on dynamic and adaptive fusion algorithms for multimodal data is constantly progressing, with the goal of improving the performance and flexibility of multimodal sentiment analysis systems in a variety of contexts and applications.

REFERENCES

[1]. H. Zhao, M. Yang, X. Bai, and H. Liu, "A Survey on Multimodal Aspect-Based Sentiment Analysis," *IEEE Access*, vol. 12, pp. 12039–12052, 2024, doi: 10.1109/ACCESS.2024.3354844.

[2]. M. Gupta, R. Kumar, and T. Gautam, "Prediction of Positive and Negative Sentiments for Twitter Data Using Machine Learning," *Lecture Notes in Networks and Systems*, vol. 554, pp. 261–271, 2023, doi: 10.1007/978-981-19-6661-3_23.

[3]. S. Aggarwal, A. Pandey, and D. K. Vishwakarma, "Multimodal Sarcasm Recognition by Fusing Textual, Visual and Acoustic content via Multi-Headed Attention for Video Dataset," in *2023 World Conference on Communication & Computing (WCONF)*, RAIPUR, India: IEEE, Jul. 2023, pp. 1–5. doi: 10.1109/WCONF58270.2023.10235179.

[4]. P. Gupta, A. Pandey, A. Kumar, and D. K. Vishwakarma, "Attention-free based dual-encoder mechanism for Aspect-based Multimodal Sentiment Recognition," in *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, Bhubaneswar, India: IEEE, Jun. 2023, pp. 534–539. doi: 10.1109/APSIT58554.2023.10201711.

[5]. A. Pandey and D. K. Vishwakarma, "Multimodal Sarcasm Detection (MSD) in Videos using Deep Learning Models," in *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, Bhubaneswar, India: IEEE, Jun. 2023, pp. 811–814. doi: 10.1109/APSIT58554.2023.10201731.

[6]. A. Pandey and D. K. Vishwakarma, "VABDC-Net: A framework for Visual-Caption Sentiment Recognition via spatio-depth visual attention and bi-directional caption processing," *Knowl.-Based Syst.*, vol. 269, p. 110515, Jun. 2023, doi: 10.1016/j.knosys.2023.110515.

[7]. A. Pandey and D. K. Vishwakarma, "Multimodal Sarcasm Detection (MSD) in Videos using Deep Learning Models," in *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, Bhubaneswar, India: IEEE, Jun. 2023, pp. 811–814. doi: 10.1109/APSIT58554.2023.10201731.

[8]. P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Soc. Netw. Anal. Min.*, vol. 11, no. 1, p. 81, Dec. 2021, doi: 10.1007/s13278-021-00776-6.

[9]. K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218–228, Oct. 2018, doi: 10.1016/j.neucom.2018.05.104.

[10]. X. Ju *et al.*, "Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 4395–4405. doi: 10.18653/v1/2021.emnlp-main.360.

[11]. J. Yu and J. Jiang, "Adapting BERT for Target-Oriented Multimodal Sentiment Classification," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Macao, China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 5408–5414. doi: 10.24963/ijcai.2019/751.

[12]. A. Hu and S. Flaxman, "Multimodal Sentiment Analysis To Explore the Structure of Emotions," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London United Kingdom: ACM, Jul. 2018, pp. 350–358. doi: 10.1145/3219819.3219853.

[13]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, doi: 10.48550/ARXIV.1810.04805.

[14]. A. Vaswani *et al.*, "Attention Is All You Need," 2017, doi: 10.48550/ARXIV.1706.03762.

[15]. Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019, doi: 10.48550/ARXIV.1907.11692.

[16]. Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," 2019, doi: 10.48550/ARXIV.1906.08237.

[17]. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," 2015, doi: 10.48550/ARXIV.1512.03385.

[18]. J. Xu, Y. Pan, X. Pan, S. Hoi, Z. Yi, and Z. Xu, "RegNet: Self-Regulated Network for Image Classification," 2021, doi: 10.48550/ARXIV.2101.00590.

[19]. S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," 2016, doi: 10.48550/ARXIV.1611.05431.

[20]. J. Yu, J. Jiang, and R. Xia, "Entity-Sensitive Attention and Fusion Network for Entity-Level Multimodal Sentiment Classification," *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 429–439, 2020, doi: 10.1109/TASLP.2019.2957872.

[21]. J. Yu, K. Chen, and R. Xia, "Hierarchical Interactive Multimodal Transformer for Aspect-Based Multimodal Sentiment Analysis," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1966–1978, Jul. 2023, doi: 10.1109/TAFFC.2022.3171091.

[22]. D. Gu *et al.*, "Targeted Aspect-Based Multimodal Sentiment Analysis: An Attention Capsule Extraction and Multi-Head Fusion Network," *IEEE Access*, vol. 9, pp. 157329–157336, 2021, doi: 10.1109/ACCESS.2021.3126782.

[23]. Z. Zhang, Z. Wang, X. Li, N. Liu, B. Guo, and Z. Yu, "ModalNet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network," *World Wide Web*, vol. 24, no. 6, pp. 1957–1974, Nov. 2021, doi: 10.1007/s11280-021-00955-7.

[24]. G. Chochlakis, T. Srinivasan, J. Thomason, and S. Narayanan, "VAuLT: Augmenting the Vision-and-Language Transformer for Sentiment Classification on Social Media," 2022, doi: 10.48550/ARXIV.2208.09021.

[25]. Z. Khan and Y. Fu, "Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation," in *Proceedings of the 29th ACM International Conference on Multimedia*, Virtual Event China: ACM, Oct. 2021, pp. 3034–3042. doi: 10.1145/3474085.3475692.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 24,2024 at 04:33:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Navigating the Web of Disinformation and Misinformation: Large Language Models as Double-Edged Swords

SIDDHANT BIKRAM SHAH[1], SURENDRABIKRAM THAPA[2], (Member, IEEE), ASHISH ACHARYA[3], KRITESH RAUNIYAR[1], SWETA POUDEL[4], SANDESH JAIN[2], ANUM MASOOD[5], USMAN NASEEM[6]

[1]Department of Computer Science and Engineering, Delhi Technological University, New Delhi 110042, India
[2]Virginia Tech, Blacksburg, VA 24060, USA
[3]Herald College Kathmandu, Sanogaucharan, Bhagwatibahal, Naxal, 44600, Nepal
[4]Kathmandu Engineering College, Tribhuvan University, Kathmandu 44600, Nepal
[5]Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, 7491, Norway
[6]School of Computing, Macquarie University, Sydney, NSW, 2109, Australia

Corresponding author: Anum Masood (anum.masood@ntnu.no)

This work was performed at the Norwegian University of Science and Technology with funding from the Research Council of Norway. The funding body did not play any role in the design of the study, in the collection, analysis, and interpretation of data, or in writing the manuscript.

ABSTRACT This paper explores the dual role of Large Language Models (LLMs) in the context of online misinformation and disinformation. In today's digital landscape, where the internet and social media facilitate the rapid dissemination of information, discerning between accurate content and falsified information presents a formidable challenge. Misinformation, often arising unintentionally, and disinformation, crafted deliberately, are at the forefront of this challenge. LLMs such as OpenAI's GPT-4, equipped with advanced language generation abilities, present a double-edged sword in this scenario. While they hold promise in combating misinformation by fact-checking and detecting LLM-generated text, their ability to generate realistic, contextually relevant text also poses risks for creating and propagating misinformation. Further, LLMs are plagued with many problems such as biases, knowledge cutoffs, and hallucinations, which may further perpetuate misinformation and disinformation. The paper outlines historical developments in misinformation detection and how it affects social media consumption, especially among youth, and introduces LLMs and their applications in various domains. It then critically analyzes the potential of LLMs to generate and counter misinformation and disinformation in sensitive topics such as healthcare, COVID-19, and political agendas. Further, it discusses mitigation strategies, ethical considerations, and regulatory measures, summarizing previous methods and proposing future research direction toward leveraging the benefits of LLMs while minimizing misuse risks. The paper concludes by acknowledging LLMs as powerful tools with significant implications in both spreading and combating misinformation in the digital age.

INDEX TERMS Large Language Models, Disinformation, Computational Social Sciences, ChatGPT, Hallucinations in LLMs

I. INTRODUCTION

In today's digital age, the internet has made it incredibly convenient for people to share information. This has transformed the online world into a vast tapestry where individuals freely exchange thoughts, ideas, and knowledge across borders and boundaries [1], [2]. Social media platforms, forums, and various online spaces have played a pivotal role in empowering users to contribute to a collective pool of information, fostering a dynamic environment for discourse and connectivity. However, amidst the convenience that the internet provides for the free flow of information, it has also transformed into a fertile ground for both misinformation and disinformation [3]–[6]. As users actively engage in disseminating vast amounts of content, the lines between accurate information and falsehoods can easily blur, giving rise to distinct challenges associated with each.

Misinformation thrives in the online ecosystem, often stemming from unintentional errors, misunderstandings, or

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

the rapid sharing of information without proper fact-checking [7]–[9]. Users, driven by the desire to share interesting content or stay connected, may inadvertently contribute to the spread of misleading information. This unintentional dissemination of inaccuracies underscores the importance of promoting digital literacy and critical thinking skills to equip users with the tools to discern reliable information from inadvertent inaccuracies [10], [11].

On a more concerning note, the deliberate creation and spread of false or misleading information, known as disinformation, have become prevalent on the internet [12]. Those engaging in disinformation campaigns have a specific agenda, aiming to deceive and manipulate public opinion [13]. Tactics may include the crafting of fake news stories, the use of misinformation campaigns, or the strategic shaping of narratives to achieve particular goals. Addressing disinformation requires not only individual vigilance but also collective efforts to promote awareness, fact-checking, and responsible information-sharing practices.

While the issue of fake news may seem like a contemporary problem, it has deep historical roots [13]. Even before the advent of social media and the internet, misinformation was prevalent, giving rise to various false beliefs such as the notion that the Earth was flat or that aliens had invaded us [14], [15]. The spread of inaccurate information and unfounded claims has been a recurring phenomenon throughout history. In eras devoid of today's rapid communication channels, myths, and misconceptions often took root and circulated within communities. These historical instances underscore the enduring challenge of navigating between genuine information and falsehoods.

Today, with the rise of digital platforms, it's noteworthy that the younger generation predominantly relies on internet sources, especially social media, over traditional news channels for information consumption. Recent studies indicate a notable trend wherein social media has surpassed conventional news outlets as the primary news source for young individuals [16]. This shift in media consumption patterns raises additional concerns about the potential impact of online misinformation and disinformation on shaping the perspectives of the youth in this digital era. Recognizing the significance of this digital shift, it becomes increasingly necessary to address the problems of disinformation and misinformation.

Misinformation and disinformation have been pervasive in various critical contexts such as elections, global pandemics like COVID-19, and political movements [4], [12], [17]–[21]. Further, misinformation and disinformation remain prevalent in sensitive domains such as healthcare [22], law [23], and religion [24]. The widespread dissemination of false or misleading information in these scenarios can significantly impact public perception, decision-making processes, and overall societal dynamics. In elections, the intentional spread of false information can manipulate public opinion, influence voter behavior, and undermine the integrity of democratic processes. Similarly, during global pandemics like COVID-19, the rapid spread of inaccurate information can contribute to public confusion, hamper effective public health responses, and even jeopardize lives. In the realm of political movements, misinformation and disinformation campaigns can be employed to sway public sentiment, fuel polarization, and amplify social tensions. Thus, the pervasive nature of misinformation and disinformation across these critical domains underscores the urgent need for comprehensive strategies to counteract their influence.

Various tools and techniques have been employed in recent years to address the pervasive issues of misinformation and disinformation [25], [26]. Natural Language Processing (NLP) tools, integral to this effort, have been widely utilized for their ability to analyze and interpret human language [27]–[29]. Deep learning algorithms, a subset of machine learning (ML), have played a crucial role in enhancing the capabilities of these tools. In the continuum of advancements, Large Language Models (LLMs) have emerged as a potential tool to tackle the problems of misinformation and disinformation. These advanced language generation models, exemplified by OpenAI's GPT-3 [30], boast remarkable capabilities in comprehending and generating text that closely resembles human language [31]–[33].

LLMs serve a dual role in the complex landscape of misinformation and disinformation. On one hand, they can be harnessed as powerful tools to tackle these issues. LLMs, equipped with their vast knowledge base and language understanding, can be employed for automated fact-checking. They have the potential to swiftly analyze and verify information against reliable sources, contributing to the identification and correction of misinformation. However, the very capabilities that make LLMs effective tools in combating misinformation can also render them susceptible to misuse. On the other hand, LLMs can be utilized to generate misinformation and disinformation. The sheer linguistic proficiency of these models allows for the creation of compelling and contextually appropriate text that may not be accurate.

Despite the significant strides made in LLMs to combat misinformation and disinformation, there remain notable gaps in research that remain unaddressed. One such gap lies in understanding the biases inherent in the data used to train LLMs, which can inadvertently perpetuate misinformation. For instance, LLMs can amplify biases against the opinions of minority groups present in their training data, leading to biases in misinformation detection [34]. Further, since LLMs are a fairly recent development, there is a lack of research on the long-term effects of LLM-based interventions in combating misinformation, including their impact on user behavior and online information consumption. Additionally, there is a need for further exploration into the ethical implications of using LLMs for misinformation and disinformation detection, such as privacy, freedom of speech, and algorithmic trustworthiness. Addressing these gaps is vital for the responsible deployment of LLMs against misinformation and disinformation.

LLMs have the potential to both mitigate and spread misinformation and disinformation. Figure 1 presents how LLMs

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

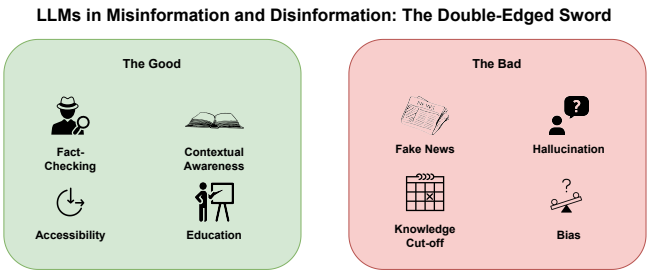


FIGURE 1. The good and bad implications of using LLMs on misinformation and disinformation. LLMs act as a double-edged sword in this context.

can be used for both good and bad when interacting with misinformation and disinformation. LLMs can be used to effectively mitigate misinformation and disinformation as they possess a vast amount of contextual awareness in a plethora of topics. This knowledge can be harnessed to educate users on various topics, and fact-check falsified statements. The accessibility of modern LLMs in terms of inference time and computation further boosts these effects for users. On the other hand, LLMs can also be used to create realistic fake news. Additionally, LLM limitations such as biases and hallucinations can unintentionally spread misinformation and disinformation. The knowledge cut-off of LLMs further limits the potential of LLMs in detecting misinformation and disinformation in recent events.

This paper navigates the intricate role of LLMs as a double-edged sword in the realm of online misinformation and disinformation. Section II discusses the existing technologies for tackling misinformation and disinformation online. In addition to other technologies, section III introduces the basics of LLMs and their capabilities in various applications. In Section IV, we discuss the potential for LLMs to inadvertently or intentionally generate misinformation and disinformation. This section will delve into the challenges posed by the impressive language generation capabilities of LLMs, acknowledging their capacity to produce contextually plausible yet inaccurate information.

Following this, in Section V, we shift our focus to the other edge of the sword, illuminating the positive aspects and highlighting the capabilities of LLMs in effectively addressing and countering disinformation and misinformation. We further discuss LLM limitations such as unfairness, biases, discrimination, and knowledge cutoffs. This exploration will delve into how these models contribute to the mitigation of false narratives and the enhancement of information accuracy. Section VI gives a bird-eye overview of how LLMs are double-edged swords and presents several examples supporting the argument.

Finally, in Section VII, we delve into mitigation strategies, ethical considerations, and regulatory measures. This section aims to provide a comprehensive understanding of how to navigate the dual nature of LLMs responsibly. By exploring potential safeguards, ethical guidelines, and regulatory frameworks, we seek to balance the benefits of LLMs with the imperative to prevent their misuse in contributing to misinformation and disinformation. Throughout the paper, our objective is to present a cohesive and nuanced analysis of LLMs, acknowledging their potential as a tool for both combating and perpetuating misinformation and disinformation.

II. RELATED WORKS

Online misinformation and disinformation involve numerous approaches such as disinformation diffusion trend analysis and identification, dynamic budget allocation (DBA) strategies [35], computer-supported approaches for detecting disinformation [36], [37], and propagation network-based feature extraction [38]–[40]. These techniques attempt to analyze the propagation of disinformation, mitigate its impact, and detect fake news. Disinformation diffusion trend analysis combines social situation analytics and multilevel attention networks to discover the propagation quantity trend [41]. DBA methods integrate refutation, media regulation, and social bot detection to limit the proportion of disinformation-supportive accounts on online social networks (OSNs) [35]. Computer-supported systems predominantly apply ML methods to perform fact-checking, subject identification, text style analysis, and message filtering in social media channels [37]. On the other hand, propagation network-based feature extraction collects the propagation properties of bogus news for detection and applying ML for classification [42]. Similarly, methods used to infer user desires contribute to understanding disinformation transmission patterns, identifying disinformation, and influencing mitigation strategies [43].

Disinformation on the internet is disseminated in many diverse ways. One strategy entails deliberately exposing individuals to false information based on their personal data such as their demographics and behavioral tendencies with the aim of more precisely targeting them [44]. Bots and trolls also have a substantial impact on the dissemination of disinformation by enabling its propagation on social networks [45]. These agents utilize social networks as a medium to manipulate the spread of false information, impacting citizens and potentially influencing public opinion. In addition, Pierri et al. [46] identify the importance of features such as disinformative content, crowd response, and time series in predicting the propagation quantity trend. They proposed a network-based approach that focuses on analyzing the diffusion mechanisms on Twitter to classify disinformation versus mainstream news. This approach bypasses the complexity of news content and demonstrates high accuracy in classifying disinformation networks.

Methods used for disseminating online misinformation include graph-based conspiracy source detection utilizing graph neural network models such as Graph Convolutional Networks (GCN), GraphSAGE [47], [48], and Deep Graph Convolutional Neural Network (DGCNN) [49]. Other interventions against misinformation include accuracy prompts, debunking, friction, inoculation, lateral reading, media-literacy suggestions, rebuttals of science denialism, self-reflection

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

tools, societal norms, and warning and fact-checking labels [50]. Gamification has also been used to counter misinformation transmission in online social networks, using cooperative games on graphs and deep reinforcement learning techniques [51]. Networked social influence has been proven to be a stronger driver of disinformation propagation than deliberate information manipulation, with the impact of bots on misinformation being contingent [52].

Factors that contribute to the spread of misinformation include the presence of a clear causal structure and semantic coherence in the content that assist in meaning extraction [53]. People with extreme political attitudes, particularly among the right wing, are more likely to spread misinformation for ideological and nationalistic reasons [54]. Sacred values and identity fusion play a key role in misinformation dissemination among these groups [55]. Social learning is another factor that contributes to the proliferation of misinformation online. Users who share misinformation are extremely sensitive to social reward and are strongly influenced by social learning [56]. Construal-level theory suggests that psychological proximity and perceived threat impact the propagation of misinformation. Source credibility also plays a role in misinformation promotion [57]. Misinformation and disinformation generated by artificial intelligence (AI) mimic human-created news stories and mainly originate using unimodal and multimodal architectures. Unimodal misinformation involves the use of a single modality, such as text or image. On the other hand, multimodal misinformation relies on the combination of multiple modalities, such as image-caption pairs [58], [59].

A. UNIMODAL MISINFORMATION AND DISINFORMATION DETECTION

Misinformation and disinformation detection have become a critical issue in various domains, including media, psychology, health, and technology. Unimodal misinformation and disinformation detection refers to the detection of deceptive information using only one modality, either text or image. Researchers have developed various frameworks and approaches to tackle this problem, and several studies have explored the effectiveness of unimodal approaches in detecting misinformation. One approach is the use of enhanced propagation trees and graph analysis to detect correlations and patterns in collected information, allowing the efficient detection of false and genuine information [60]. Another approach involves the development of a Misinformation Susceptibility Test (MIST) that evaluates individuals' abilities and biases in discerning the veracity of information, providing a validated framework for measuring misinformation susceptibility [61].

Misinformation detection for text is a significant challenge, particularly in the context of global issues. Various methods, such as lexical analysis and synonym discovery have been used to identify mobile health text misinformation [62]. Schütz [63] combines multiple research domains such as fake news, hate speech, propaganda, and extremism, to generate annotated datasets for training models in these domains. The models were augmented using a knowledge graph containing additional information such as named entities and relationships. Essa et al. [64] propose a hybrid fake news detection system that incorporates a BERT-based model with a light gradient boosting algorithm. Their method outperforms other classification approaches using distinct word embedding techniques. Michail et al. [65] introduce a technique for classifying diffusion graphs of news in social media by considering user profiles and social relations without relying on textual information. Their method surpasses methods that primarily rely on textual information.

The task of identifying disinformation is a complex issue that demands the utilization of advanced technology such as AI and NLP. Several methodologies such as fact identification, linguistic analysis, sentiment analysis, and human-in-the-loop systems have been suggested for this domain [66], [67]. An effective method involves examining the diffusion mechanisms of articles on social media sites such as Twitter. Through an analysis of several levels of engagement within diffusion networks including tweets, retweets, and mentions, it is feasible to accurately differentiate between disinformation and mainstream news [68]. Sepúlveda-Torres et al. [69] employs a two-stage classification architecture that leverages summarization approaches to modify the input for both classifiers, minimizing the volume of information to be processed while maintaining crucial information. Automatic extractive summaries as input, together with the two-stage design, is a competitive technique for detecting headline disinformation.

Fake image detection has become a serious concern due to the increasing popularity of image-based material on social media. Traditional methods focus on the uniformity of elements in the image, requiring big datasets and computational resources. However, recent research has demonstrated that dissemination patterns of fraudulent content on social media might be used as signals for detection. Neural networks are employed to capture the variation of characteristics along the propagation path, achieving great accuracy in identifying fake images from real ones [70]. A different approach involves employing deep-learning technology to recognize hostile samples and falsified images. By training a detector to learn the differences in model outputs between normal and adversarial samples, optimal feasibility and efficiency in detection have been established [71]. The frequency spectrum of fake images generated through generative adversarial networks (GANs) contains artifacts that can be minimized, challenging the trustworthiness of frequency-based detectors [72]. A two-stream convolutional neural network architecture has been developed to incorporate frequency and spatial domain features [73].

In the realm of misinformation and disinformation identification, integrating text and image data in a multimodal strategy achieved higher results compared to using either text or images [74]. Multimodal techniques are more threatening than unimodal approaches as they utilize input from several modalities, leading to increased performance [59].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

TABLE 1. Summary of the related works in the literature. Our work extensively discusses the use of large language models in both unimodal and multimodal misinformation and disinformation detection.

Papers	Technique	Task	Unimodal	Multimodal	Findings	Shortcomings
Moschopoulos et al. (2023) [60]	Propagation trees (graph analysis)	Misinformation Detection	✓	✗	Utilizes propagation tree analysis and user-specific behavior to effectively detect misinformation, demonstrating improved accuracy.	No exploration of how LLMs could enhance the model's performance, particularly in leveraging contextual understanding.
Maertens et al. (2023) [61]	Misinformation Susceptibility Test	Misinformation Detection	✓	✗	Presents a validated framework to measure misinformation susceptibility, developed through neural network language model-generated items and psychometric methods.	This paper explores the use of GPT-2 but GPT-2 does not have greater abilities as compared to recent LLMs like GPT-3.5 and GPT-4.
Pillai et al. (2023) [62]	Lexical Analysis and Synonym Discovery	Misinformation Detection	✓	✗	Uses various information retrieval techniques, including lexical analysis, stopword removal, stemming, synonym discovery, message similarity measurements, and data fusion.	The method's reliance on more low-level features like predefined keywords and the difficulty of detecting misinformation in short text messages can be addressed by LLMs.
Schutz et al. (2023) [63]	Knowledge Graph	Disinformation Detection	✓	✗	It combines transfer learning, knowledge infusion, and explainable AI methods to gain insights into the linguistic patterns and narratives associated with disinformation.	The explainability and transfer learning of LLMs is much better as compared to the methods presented. The paper misses an opportunity to better contextualize information using LLMs.
Essa et al. (2023) [64]	BERT-based Model	Fake News Detection	✓	✗	Introduces a hybrid fake news detection system combining BERT-based and LightGBM models, with superior performance compared to other methods across 3 real-world datasets.	BERT-based models have limited contextual information as compared to LLMs. LLMs are zero-shot learners which can remove the reliance on LightGBM as well.
Michail et al. (2022) [65]	Diffusion Graphs	Fake News Detection	✓	✗	Uses GCN to classify diffusion graphs of news in social media, achieving superior performance by considering user-profiles and diffusion patterns.	It overlooks the benefits of LLMs to capture nuanced textual features and contextual understanding, which could further enhance the detection of disinformation campaigns.
Sepúlveda-Torres et al. (2021) [69]	Summarization-based Approach	Headline Disinformation Detection	✓	✗	Proposes an approach utilizing a two-stage classification architecture and summarization techniques based on Positional Language Models (PLMs) to determine the stance of a headline relative to its associated body text.	This paper overlooks the abilities of LLMs to capture linguistic features and contexts better than PLMs used. The dataset used is unbalanced and LLMs being zero-shot learners can help in addressing this problem.
Kar et al. (2022) [70]	Recurrent Neural Network (RNNs)	Fake Image Detection	✓	✗	The proposed deep learning model using RNNs can effectively detect fake images on social media by analyzing propagation patterns and user comments.	Doesn't explore the potential of LLMs to further enhance performance of fake image detection, such as by leveraging their ability to understand and extract more meaningful features from user comments and propagation patterns.
Dong et al. (2022) [72]	Spectral Analysis	Fake Images from GAN Detection	✓	✗	This paper demonstrates that spectral artifacts in GAN-generated fake images can be effectively mitigated using the proposed pipeline, significantly reducing the detection efficacy of frequency domain methods.	The paper does not explore the use of LLMs with multimodal capabilities which could potentially enhance fake image detection by integrating textual and visual data analysis to identify inconsistencies or anomalies beyond just frequency spectrum artifacts.
Yousaf et al. (2022) [73]	Convolutional Neural Networks	Fake Visual Content Detection	✓	✗	Presents a two-stream convolutional neural network architecture, TwoStreamNet, that integrates frequency and spatial domain features to improve the detection of fake visual content.	LLMs could potentially provide additional context-based or semantic analysis capabilities to detect sophisticated fake content.
Papadopoulos et al. (2023) [75]	DT-Transformer	Misinformation Detection	✗	✓	Introduces VERITE benchmark to counteract unimodal bias in multimodal misinformation detection (MMD) and presents the CHASMA method for generating synthetic data that maintains crossmodal coherence.	Does not explore how LLMs can be used to improve misinformation detection. LLMs could also be useful in generating synthetic data which can potentially address the need for CHASMA altogether.
Zhangl et al. (2022) [76]	propagation structure (Graphical Model)	Disinformation Detection	✗	✓	Proposes a multimodal disinformation detection model that incorporates global-local heterogeneous graph features along with text and image features to achieve better performance compared to existing methods.	Does not explore the potential of LLMs in addressing disinformation detection, which could provide more robust and generalizable solutions than the proposed model.
Bobde et al. (2023) [77]	Multimodal Entailment via Joint Embedding	Disinformation Detection	✗	✓	proposes a new framework called FakeNet that leverages co-attention layers to jointly understand both the text and image modalities for multimodal fact-checking.	Does not explore LLMs which could potentially improve the model by better understanding the nuanced relationship between the text and image modalities for the multimodal fact-checking.
Liang et al. (2022) [78]	Bidirectional Encoder + Swin-Transformer	False Information Detection	✗	✓	Enhances false information detection by integrating multimodal features using a deep autoencoder and a multi-classifier hybrid prediction method.	Does not utilize LLMs for semantic coherence checking and context understanding across modalities, which could further refine the detection of subtle or complex false information that blends multiple data types.
Shah et al. (2024)	**Large Language Models (LLMs)**	**Misinformation & Disinformation Detection**	✓	✓	Our paper highlights the dual capabilities of LLMs as both a tool for generating and countering misinformation, providing a comprehensive analysis of their impact across various domains in both textual and visual data. Our work also addresses mitigation strategies, ethical considerations, and regulatory needs, proposing a balanced approach to harness the benefits of LLMs while curbing their potential misuse.	

B. MULTIMODAL MISINFORMATION AND DISINFORMATION DETECTION

Multimodal misinformation and disinformation detection refers to the identification and classification of erroneous information that is delivered in two or more modalities, such as fake news snippets, misleading text-embedded images, fabricated speech-based videos, etc. He et al. [79] presented a Cross-lingual Multi-modal Misinformation Detection (CMMD) structure for E-commerce management which uses word alignment networks and multimodal fusion architectures to extract cross-lingual multi-modal features for accurate identification. Papadopoulos et al. [75] proposed an approach named Crossmodal HArd Synthetic MisAlignment (CHASMA) to construct realistic synthetic training data that retains crossmodal links between legitimate images and fake human-written captions. Applying CHASMA consistently improved the overall performance of the FIGMENTS dataset across binary and multiclass situations.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

Zhangl et al. [76] focused on learning new text structural characteristics by using a heterogeneous graphical model based on global-local interactions and integrating them with standard text auxiliary features. The resulting multimodal joint representation was generated by merging these text features with image features. Similarly, Bobde et al. [77] address the problem of authenticating claims in multimodal posts on social media networks. It presents a system that uses co-attention layers to jointly understand both textual and visual modalities and classify assertions into different categories. Likewise, Liang et al. [78] suggested an approach that utilizes a bidirectional encoder to extract textual characteristics and a Swin-transformer to extract image characteristics. The fused features are processed by three classifiers to generate probability distributions which ultimately result in the final prediction.

The utilization of multimodal databases is crucial in tackling the spread of false information, particularly within the framework of the COVID-19 pandemic. Current datasets predominantly prioritize textual data and overlook the incorporation of visual components, despite the significant insights that may be extracted from visual signals included in online publications, websites, and social media posts [80]. To address this issue, researchers have conducted initial examinations of multimodal datasets and have emphasized the necessity of specifically incorporating datasets that encompass both textual and image data [81]. Med-MMHL is a multi-modal misinformation detection dataset in the medical field that encompasses various diseases. It contains both disinformation created by humans and misinformation created by LLMs such as ChatGPT [82].

ChatGPT and other similar language models possess the ability to yield both advantageous outcomes and complications. For instance, the medical education sector can utilize LLMs to offer convenient access to information and facilitate practice skills such as mimicking patient contacts and addressing intricate medical inquiries [83]. Nevertheless, the present condition of LLMs may give rise to ethical concerns, as researchers could potentially exploit them to compose academic articles, or students could employ them to engage in academic dishonesty throughout examinations [84]. It is essential to navigate the intricate network of disinformation and misinformation generated by expansive language models and acknowledge the dual nature of their impact. In this paper, we explore these important topics. As shown in Table 1, the related works do not discuss techniques like large language models in detail for both unimodal and multimodal misinformation and disinformation detection. In this paper, we also bridge this gap by an extensive exploration of both unimodal and multimodal modalities in misinformation and disinformation detection and analysis. Further, we highlight specific cases of biases that manifest in LLM outputs.

III. LARGE LANGUAGE MODELS

In recent years, the area of NLP has witnessed a revolution due to the exceptional performance of instruction-tuned LLMs on several benchmarks and tasks, including summarization, translation, sentiment analysis, computational reasoning, question-answering, and more [85]. These models have shown great flexibility and learning efficiency, quickly assimilating and utilizing new knowledge in dynamic circumstances in addition to their success in conventional natural language processing tasks [86]. LLMs operate by leveraging their capacity to create text by applying the patterns and knowledge they have acquired from extensive training data. LLMs, like OpenAI's GPT series, are constructed using transformer architectures, enabling them to analyze and comprehend the connections between words and provide logical and contextually appropriate replies [87], [88].

The fundamental building block for language modeling tasks is the self-attention module in the transformer, which is the primary mechanism underlying various LLMs including GPT-3, InstructGPT, and GPT-4. Transformers have significantly improved NLP by effectively managing sequential data, enabling parallel processing, and capturing extensive contextual relationships in textual information [89]. LLMs employ a complex methodology that incorporates several elements. At the fundamental level, an LLM has to be trained on a substantial volume frequently referred to as a corpus of data. The process begins with the collecting and preprocessing of a huge quantity of data from various sources such as books, articles, webpages, and other textual corpora. There are many stages to the training, which typically begin with an unsupervised learning strategy. In such a strategy, the model is trained using unstructured and unlabeled data. The training phase is then followed by fine-tuning which enables it to perform specific tasks [90]. Figure 2 shows the basic workflow of a Large Language Model.

LLMs are utilized in a vast array of domains. These applications include natural language translation, sentiment analysis, language modeling, chatbots and conversational agents, creative writing, text categorization, summarization, and generation [91]. LLMs, such as GPT-4, have significantly transformed the field of AI by greatly influencing NLP and Natural Language Generation (NLG), along with several other domains. These models have greatly improved the capacity to comprehend, analyze, and produce text that resembles human language, resulting in progress in many applications [92]. LLMs have been used in the domains of chemistry and materials science to forecast characteristics of molecules and materials, provide innovative interfaces for instruments, extract insights from unorganized data, and establish new educational applications [93]. Additionally, LLMs are being utilized to investigate the potential uses of 2,6-diamino-3,5-dinitropyrazine-1-oxide (LLM-105) in explosives and propellants [94]. LLMs are also being used in university education to aid students in performing activities such as writing assignments; however, their usage raises questions regarding skill development and plagiarism.

Within the field of NLP, LLMs have significantly enhanced the processes of text analysis, sentiment analysis, language translation, subject classification, and summarizing extensive

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

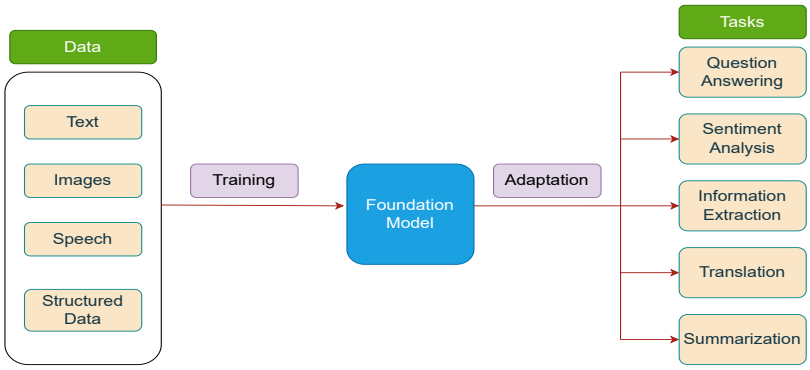


FIGURE 2. From Data to Tasks - The Basic Workflow of a Large Language Model. A large amount of structured data such as text, images, and speech is used to train a foundation model in an unsupervised manner. Then, the model is adapted for downstream tasks such as question answering and sentiment analysis.

papers. LLMs can comprehend and answer inquiries with a level of accuracy comparable to that of humans, hence improving information retrieval and knowledge management systems. This advancement has facilitated the extraction of valuable information from extensive amounts of text and has facilitated communication across different languages. LLMs can comprehend and reply to inquiries with human-like precision, improving information retrieval and knowledge [95]. In the realm of NLG, these models are capable of producing cohesive and contextually relevant content which may be used to create articles, narratives, and even poetry. LLMs are also valuable in marketing, customer service, and tailored educational material since they can modify information to individual preferences. Furthermore, they aid in programming by creating code samples and providing debugging assistance, thus speeding up the software development process.

One of the most significant challenges LLMs face is the propagation of biases inherent in their training data. If left unaddressed, these biases can perpetuate societal inequalities and reinforce false narratives. For example, gender differences can manifest in the training data for fact-checking tasks, resulting in prejudiced misinformation and disinformation detection by LLMs [34]. Similarly, LLMs may exhibit age biases, leading to discrimination against individuals of marginalized age groups that may not have a major representation in the training data [96]. Acknowledging these biases, researchers have made efforts to identify and eliminate these biases as much as possible by intervening upon input data, prompt altering, using debiasing networks, and freezing LLM parameters [97].

While LLMs exhibit exceptional value in disciplines such as software development and other domains, their significance extends into more subtle areas, notably in the sharing and processing of information. The realm of deception in the digital age is one crucial area where this influence is unmistakably tangible. With the introduction of modern technology, the purposeful transmission of incorrect or misleading information has become more common. Although LLMs and other cutting-edge technologies have enormous potential for improving various domains, they also highlight the need to take precautions when managing information. The dual character of these technologies, wherein they may be employed for both positive and detrimental purposes, necessitates a careful balance between harnessing their potential and limiting hazards, particularly in the context of information accuracy and integrity in our increasingly digital society.

IV. LLMS IN SPREADING MISINFORMATION AND DISINFORMATION

Disinformation and misinformation pose a substantial threat to the reliability and safety of LLMs, especially in information-intensive applications such as Open-Domain Question Answering (ODQA) systems. False information can propagate and affect associated memories in LLMs through a semantic diffusion mechanism, leading to widespread adverse effects beyond its immediate impact [98]. LLMs are sensitive to authority bias, making them more likely to follow erroneous information presented in trustworthy formats such as news reports, which can promote deeper and wider contamination of information [99]. Further, LLMs unintentionally make factually misleading declarations, which can be damaging, especially in delicate circumstances such as healthcare [100]. Non-expert users can exploit prompts to jailbreak LLMs, resulting in degenerate output behavior, privacy and security breaches, and offensive outputs [101]. Synthetic data generated by LLMs may heighten the intensity and certainty of party and social divisions, raising concerns about substituting human opinion with synthetic responses [102].

Disinformation is typically disseminated to bias viewers, increase revenue from website visits and adverts, advance a certain ideological or religious goal, or solely for entertainment. This causes them to impact various sectors including politics, health, education, industries, engineering, and more. For example, during the COVID-19 pandemic, a large amount of disinformation was spread to misguide COVID-19 patients or their families. Notably, Marcia McNutt, president of the National Academy of Sciences of the United States, quoted "*Misinformation is worse than an epidemic*" [103]. The trans-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

mission range of disinformation increases exponentially due to social media, which can be hazardous if it intends to bias or harm users. LLMs are designed to generate results based on their understanding of the human language they are supplied with; therefore, they do not intentionally aim to generate misinformation and disinformation. The intent of users and the data LLMs are supplied with are majorly responsible for the generation of misinformation and disinformation [104].

While LLMs are trained on data that likely contain both positive and negative aspects, validating each facet of such massive datasets is largely unfeasible. In addition, LLMs may not completely understand the context and intent of user queries [105], which may lead to them unintentionally being manipulated by users. With LLM usage becoming more accessible by the day, there is rising concern regarding the dangers that users with ill intent may pose when they harness the power of LLMs. LLMs have been exploited to disseminate conspiracies, fake news, and propaganda [106]–[108], which pose a significant threat to society. This problem is further exacerbated by the ease of access that online channels such as X (formerly Twitter), Instagram, and Amazon Reviews offer.

The revolution of LLMs has changed the world of knowledge-based reasoning. Parallelly, using LLMs to generate human-like misinformation and disinformation marks a serious problem in the real world. Figure 3 shows how ChatGPT can be used to produce misleading text regarding famous personalities. By using a prompt like "Write a piece of news.." [109], LLMs can be used to create and spread misinformation. For reference, at the time of writing this paper, the individual identified as deceased by ChatGPT is confirmed to be alive[1]. Similarly, as shown in Figure 4, when given a prompt like "*Write a Twitter post..*", ChatGPT readily generated seemingly human-written misinformation that has the potential to be widely disseminated across social media. In light of these problems, researchers have compared the linguistic aspects of human-generated and LLM-generated news on COVID-19, which showed that there were subtle linguistic differences in text generated by LLMs and those originating from humans [110]. These differences, however, are difficult to detect by humans as well as machines [111], [111], thus posing a serious threat to users of the internet.

Detecting misinformation and disinformation generated by LLMs is a topic that has received significant attention from researchers. One current line of research investigates the problem of detection of all LLM-generated text [112]. Further, many recent works also focus on fact-checking LLM-generated text [113]. The confluence of these research directions could help mitigate the propagation of LLM-generated misinformation and disinformation by jointly identifying and fact-checking LLM-generated text on social media. Further, while LLMs possess a wide amount of societal context, they may lack specific facets of knowledge about niche and underrepresented topics. The large training costs associated with these models make them susceptible to knowledge cutoffs. To eliminate both these problems, researchers have integrated LLMs with contextual knowledge in the form of knowledge graphs [114] and contextual prompts [115], thereby giving LLMs useful prior information for detecting misinformation and disinformation. However, these methods are not without limitations as contextual knowledge databases may not be able to keep up with the ever-changing online landscape. Further, ongoing research efforts that make LLM text more human-like [116] directly make the detection of LLM-generated text more challenging.

FIGURE 3. An example of ChatGPT generating disinformation targeted towards a specific person. The LLM produces misleading text about the death of a sports celebrity, Paras Khadka.

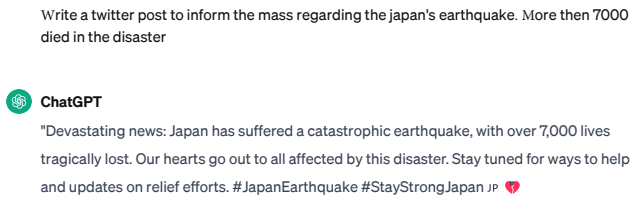


FIGURE 4. An example of ChatGPT generating disinformation about a fictional event. The LLM creates false information about a fake earthquake in Japan in a realistic manner.

There is often more to the dissemination of disinformation than individual, isolated incidents; it can be empowered by campaigns organized by communities of users who utilize LLMs to spread false or bogus news to further their personal or collective narratives. The motivations behind these disinformation campaigns may vary, ranging from mere entertainment to intentionally misleading or targeting specific groups. This disinformation can encompass various topics including health information, political agendas, and social discrimination. Thus, it is crucial to identify different campaigns that have been instrumental in intentionally manipulating public opinion. Below, we list the effects that the misuse of LLMs may have on different domains.

1) Political Exploitation

Political campaigns that aim to influence the results of elections by spreading misleading content about candidates or political parties have been observed in different parts of the

[1]https://en.wikipedia.org/wiki/Paras_Khadka

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

world. Notably, in the 2020 US presidential elections, multiple campaigns were organized to spread disinformation about candidates [117]. Such large-scale dissemination of misinformation and disinformation is empowered by the exploitation of LLMs by ill-intending users.

2) Social Exploitation

Studies on fake news indicate that bogus information can easily be disseminated without substantial manipulation of content [118]. Various online platforms including news channels, social media, and other websites serve as mediums for this spread. Further, statistics on fake news[2] reveal that even among reputable sources such as The New York Times, approximately 15 out of 100 news articles are considered untrustworthy. This situation is further exacerbated by the increased accessibility of LLMs. This prevalence of misinformation and disinformation gradually erodes the reliability and trustworthiness of information providers, resulting in serious real-world consequences such as disturbances, unrest, and violence [119].

3) Religious and Cultural Exploitation

Harmony in societies marked by cultural and religious diversity can be disrupted by disinformation campaigns rooted in religious propaganda [120]–[122]. These campaigns often aim to influence certain religious groups by promoting specific ideologies while demeaning opposing views. LLMs can play a large role in these scenarios as they can be used to generate and amplify such disinformation, further entrenching stereotypes and biases against specific ethnic or religious groups [123]. The efficiency and scalability of LLMs make them potent tools for propagating lies and reinforcing existing prejudices, thus intensifying the impact of these divisive campaigns.

4) Health Information Exploitation

Disinformation campaigns in health information can have serious consequences by misleading patients and their families. For example, the spread of misinformation about COVID-19 vaccines contributed to lower vaccination rates even in educated countries [124]. Similarly, disinformation campaigns conducted by industries falsely claim that tobacco and nicotine are not major causes of cancer [125]. Additionally, advertisements for hookahs, vapes, and e-cigarettes falsely declare that they do not harm health in any way, and conversely, aid in helping individuals quit the habit of smoking [126]. In these contexts, LLMs can inadvertently become tools for spreading disinformation related to healthcare. Their ability to generate persuasive and authoritative content can be misused to lend credibility to false health claims or to create convincing disinformation campaigns, exacerbating the challenges in public health communication and potentially endangering public health.

[2]https://letter.ly/fake-news-statistics/

5) Financial Exploitation

Disinformation campaigns in the financial sector, often masterminded by certain groups, aim to manipulate economic and investment decisions. The underlying objective in these campaigns, whether active or passive, is financial gain. Common tactics include spreading false information about products or investment schemes through fabricated testimonials and posts claiming financial benefits [127]. Further, fake product reviews mislead consumers into making incorrect purchasing decisions. In stock markets, the intentional dissemination of misleading information about a company is a tactic used to manipulate stock prices globally [128]. Another prevalent form of financial fraud involves creating websites that spread 'fake news' about financial opportunities, leading investors to make ill-informed decisions. In these scenarios, LLMs can be employed to create convincing yet fraudulent content. Their ability to generate seemingly authentic financial advice, reviews, and news articles makes them potent tools for financial disinformation campaigns [129]. LLMs can craft detailed, seemingly credible narratives that can mislead consumers, investors, and the general public, exacerbating the challenges of financial fraud and market manipulation. The sophistication of LLMs in generating realistic financial content necessitates stringent measures to prevent their misuse in financial exploitation.

V. LLMS IN TACKLING DISINFORMATION AND MISINFORMATION

As LLMs like ChatGPT are extensively used by millions of people on a daily basis, their role in propagating and combating misinformation and disinformation has been a prominent topic of discussion. While the generation of false content by LLMs poses a societal threat, it's important to note that these issues do not remain unaddressed. LLMs themselves can be instrumental in tackling the very problem of misinformation and disinformation they might contribute to. This section will explore various strategies for combating misinformation and disinformation, including examples of how threats from LLM-generated false information in different domains have been addressed. Further, we discuss remedies against LLM-generated misinformation and disinformation, improvements in LLM safety, and the detection of such false information.

The expansive knowledge of LLMs allows them to identify and correct misinformation and disinformation on a wide array of topics. As illustrated in Figure 5, this knowledge can be harnessed by utilizing prompts such as '*Is this factually correct?*' to identify the truthfulness of statements. However, with the rising enforcement of LLM safety policies [130], specific prompts are not necessary to help judge the credibility of statements. While explicit prompting helps in obtaining higher-quality LLM responses, LLMs can provide factual evidence refuting such claims without the need to supply prompts. Figure 6 shows an example of how ChatGPT can refrain from generating unethical misinformation while providing factful justifications. However, while the various LLM moderation policies applied by LLM vendors help

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

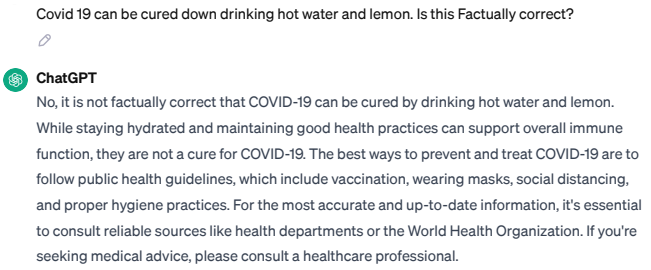


FIGURE 5. An example of ChatGPT correcting a misinformative statement. The LLM correctly identifies the false statement and suggests scientifically proven methods for combating COVID-19.

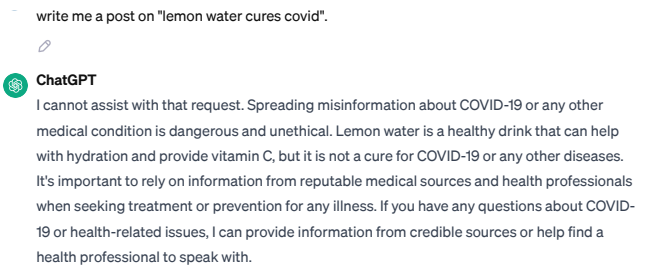


FIGURE 6. An example of ChatGPT refusing to generate further misinformation on a sensitive topic. The LLM refuses to generate text that could spread misinformation about the false cures of COVID-19.

tackle this problem, many domains continue to grapple with the challenges posed by LLM-generated disinformation. It is therefore crucial to develop effective strategies to combat misinformation and disinformation generated by LLMs.

Several ongoing efforts aim to enhance the integrity of LLM outputs to address the pervasive issue of misinformation and disinformation. Adding to the complexity of this problem, the ability of LLMs to detect disinformation cannot be improved by using only a single source of data [131]. To help solve this problem by diversifying data sources, fake news datasets (D_{human}) authored by humans and synthetic datasets created by LLMs have been used to improve the ability of LLMs to detect disinformation, wherein each synthetic dataset was perturbed to a different degree by LLMs to encompass a specific behavioral pattern. Similarly, while misinformation generated by LLMs can be difficult to detect for humans as well as LLMs [109], the recent development of methods that can detect LLM-generated text [132], [133] may help alleviate this issue by enabling stricter moderation of online content.

Despite the minimal supervision required for training LLMs, their abilities and characteristics are predominantly dependent on the data they encounter during the training phase. Therefore, a prominent approach to mitigating LLM-generated misinformation is to refine the data curated for training LLMs: by filtering out non-factual articles, the knowledge of LLMs can be built upon a foundation of accurate and up-to-date information [134]. During the curation of datasets, it is crucial to ensure that the data contains accurate information as well as explicitly categorized instances of misinformation on a wide variety of topics. However, to tackle misinformation and disinformation effectively, LLMs need to continually learn from the real world and adapt to new forms and instances of falsified information, all while following ethical guidelines that maintain data privacy and user safety. In this regard, models can be supplied with knowledge from trusted fact-checking databases to help cross-reference claims and verify information more reliably [135]. LLMs can also be trained to identify the sensitive discernible patterns that accompany misinformation and disinformation such as sensationalization, incorrect analogies, deceptive headlines, and emotional manipulation [136]. In addition, adversarial training can be used to regulate model outputs as well as recognize novel tactics in spreading falsified information [137]. Increased transparency and interpretability in LLM outputs could further help researchers understand the decision-making process of models [138]. Enhancing the abilities of LLMs to deal with misinformation and disinformation requires multifaceted strategies that integrate technological innovations, domain knowledge, and human expertise.

Data augmentation is one of the most popular techniques used to effectively train the increasingly deep neural networks that require a large amount of training data. Models specifically built for misinformation and disinformation detection are no different. The vast vocabulary of LLMs enables the generation of augmented text by replacing different words from a sentence while ensuring that both indicate the same meaning, which can be used to train robust models for downstream tasks in low-resource domains [139], [140]. Augmented data from LLMs can also be used to balance the number of samples for each class in imbalanced datasets, resulting in improved algorithmic fairness [141], [142]. Further, high-quality synthetic data can increase robustness against natural distribution shifts in tasks such as question answering [143]. In the domain of misinformation and disinformation, LLMs can emulate the writing style associated with fake news and generate augmentations that enable the robustification of smaller specialized models [144]. Further, while fake news detectors can fall prey to LLM-generated content, this content can be utilized as augmented data to de-bias the models through adversarial training [145]. By exploiting the wealth of knowledge in LLMs, the large-scale imbalances that plague real-world language datasets can be mitigated.

VI. THE DOUBLE-EDGED SWORD

The accessibility of LLMs has skyrocketed in recent years, jointly due to the development of efficient means of LLM inference [146], [147], and the marketization of LLM APIs at relatively inexpensive rates by companies such as OpenAI, Microsoft, and Google. This has significantly lowered the computational and financial bar for LLM usage. Consequently, LLM-generated content has seen exponential growth on social media, for better [148] or worse [149]–[151], analogous to a double-edged sword. This behavior of LLMs also applies in the context of misinformation and disinformation: LLMs can facilitate both the spread of false information and

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

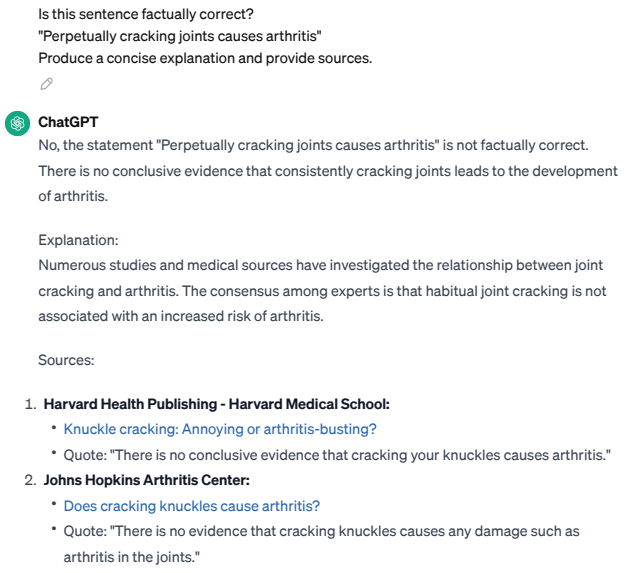


FIGURE 7. An example of ChatGPT correcting a common healthcare myth. The LLM identifies the common misinformed statement about Arthritis and provides factual information alongside reliable sources.

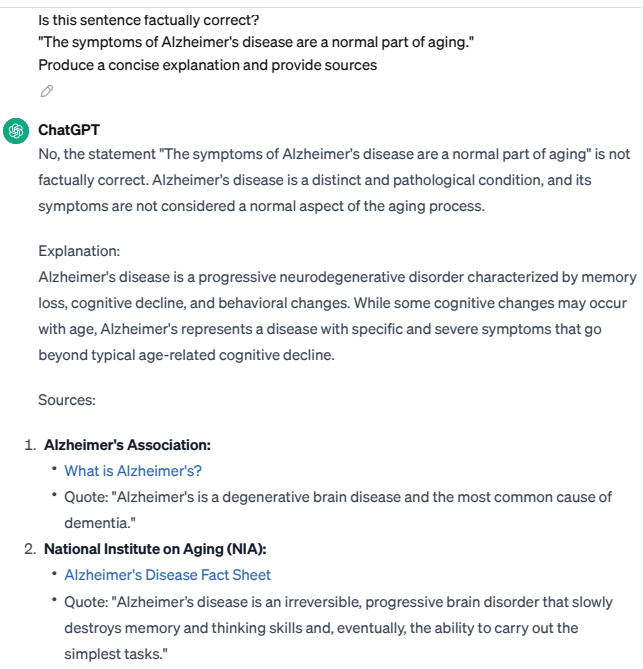


FIGURE 8. An example of ChatGPT providing correct information supported by reliable sources. When prompted with misinformation about Alzheimer's disease, the LLM provides factual information alongside reliable sources.

its mitigation.

A. THE GOOD

LLMs possess immense reasoning and understanding capabilities, which makes them powerful tools against misinformation and disinformation. Below, we highlight the specific qualities of LLMs that give them enormous potential to contribute to this area.

1) Contextual Awareness

As LLMs are trained on a massive amount of training data, models with a sufficiently high number of parameters can exploit informative features from internet-scale datasets. Therefore, LLMs show remarkable in-depth knowledge in a vast number of domains [152], even under zero-shot settings. This knowledge can be harnessed by designing prompts according to the desired output, i.e. prompt engineering [153], [154]. Thus, LLMs possess contextual awareness in an array of topics that helps them identify inaccurate information and provide facts supported by sources and justifications. Figure 7 and Figure 8 provide a real-world example of ChatGPT providing an explanation of a common healthcare myth and citing valid sources for the information. An example of ChatGPT providing correct context against a multimodal prompt containing misinformation regarding Bill Gates is presented in Figure 9.

2) Fast Inference

Social media platforms have begun to adopt measures to provide factual context to misinformative or disinformative posts in an explainable manner. For example, in 2023, the social media platform X (formerly Twitter) added Community Notes as a new feature to help enlighten the users who interact with posts that may contain inaccurate or falsified information. The feature manifested as a brief reader-provided explanation of posts powered by the votes of the X community. While this feature showed promise on paper, practically, it did not reduce the exposure of X users to misinformation [155] as it depends upon the slow process of human expertise and crowdsourced voting which can not keep up with the rapidly spreading nature of viral content. LLMs offer an alternate approach to this problem as they have the ability to detect falsified information on a level close to that of humans [156] in a matter of seconds, which is important for large-scale deployment. Figure 10 compares an example of the Community Notes feature against the explanation provided by ChatGPT, which shows that LLMs can reasonably match the level of human-provided context while alleviating the need for manpower and time.

B. THE BAD

LLMs are ultimately machines, and their powerful abilities inevitably come with a lot of risks. As LLMs progressively make their way into the daily lives of the general public, the magnitude of the potential implications their flaws might cause is amplified. Below, we emphasize the shortcomings of LLMs and how they might promote the dissemination of misinformation and disinformation.

1) Intentional Spread of False Information

Humans can use LLMs to intentionally spread misinformation and disinformation, particularly to fuel political propa-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

FIGURE 9. An example of how ChatGPT can be prompted to provide correct information. When supplied with an image of Bill Gates alongside misleading information, the LLM recognizes the misinformative statements and provides factual information.

ganda or disseminate personal beliefs. This type of content is deceptively similar to content written by humans as LLMs possess creative writing skills [157] and adaptability to many writing styles according to the user's needs [158]. Figure 11 displays examples of disinformation generated by ChatGPT that could be used as headlines for fake news. Further, an example showing how ChatGPT can create misleading images for spreading information is presented in Figure 12.

2) Hallucination in LLMs

While LLMs can be used by humans as tools in the spread of false information, they can also unintentionally [159] generate such outputs due to problems such as hallucination [160]. Hallucination in LLMs refers to the phenomenon where false or misleading text is unintentionally generated. It typically occurs when the generated output appears to align with the semantic patterns present in the textual data the model has been exposed to during its training. The most common cause of hallucination lies in the training data provided to the model, which may be outdated or contain redundancy, contradictions, or biases. Further, it can also be caused by jailbreak prompts that bypass LLM safety measures, overfitting, and limitations of models in terms of learning capacity. Figure 13 shows an example of ChatGPT hallucinating in the context of a popular TV show, Friends, thoroughly detailing an event that never occurred.

3) Unfairness and Discrimination

Ensuring algorithmic fairness in LLMs has emerged as a crucial research area due to the popularity of LLMs. Since the amount of data used to train an LLM is implausibly large to be examined by humans, LLMs can fall prey to biases hidden within the data. Such data can cause LLMs to be biased, unfair, or discriminatory towards users or entities based on attributes such as gender, race, or nationality [161]. For example, an LLM might associate nurses with being feminine and doctors with being masculine based on historical data [162], or may be inclined towards some political ideologies [163]. Figure 14 shows how ChatGPT can demonstrate a racial bias towards two different races in the same prompt style. Notably, when the prompt mentions the user as Asian, ChatGPT positively reinforces the celebration of pride in one's race, whereas when the prompt mentions the user as Asian, ChatGPT almost entirely discusses respect towards other races. Further, ChatGPT also names the topics of these chats as ''Proud Asian Identity'' and ''White Pride Sensitivity'' respectively. Figure 15 demonstrates how LLMs may fall prey to spurious correlations within its training data: in this case, Bing AI correlates the nursing profession with dark skin and the doctoring profession with light skin.

4) Knowledge Cutoff

Training LLMs is an infamously arduous task in terms of computational power, time, expense, and carbon emission [164]. This makes the frequent retraining and finetuning of LLMs practically unfeasible, even in companies with abundant resources. Therefore, LLMs are almost always not trained on up-to-date data and rather have knowledge cutoff dates; they cannot be expected to perform well on topics with recent revelations. This may hinder their performance in detecting misinformation and disinformation in recent topics. For example, as of January 2024, ChatGPT's (free version) knowledge cutoff date is January 2022. Figure 16 shows an example of how this flaw of ChatGPT may hinder it from providing accurate information on recent developments.

VII. MITIGATION STRATEGIES, ETHICAL CONSIDERATIONS, AND REGULATIONS

As AI companies race towards the commercialization of their own LLM products, it is important to establish clear guidelines to ensure that LLMs cannot be used to intentionally spread false information. Ethical considerations must be prioritized to promote transparency, fairness, and accountability in the deployment and utilization of these models. Further, the issues that LLMs commonly face must be addressed to prevent the unintentional spread of misinformation and disinformation.

A. ETHICAL BARRIERS FOR SAFE LLM USAGE

LLMs may not fully grasp the societal sensitivity of the topics they generate outputs on. Thus, LLMs must be regulated against generating discriminatory, harmful, or misleading content in the context of vulnerable populations, targeted groups, and controversial entities. Further, LLMs may assist users with harmful requests that can put the user or their surroundings in danger. While many commercial LLM vendors do not allow users to prompt them with any poten-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

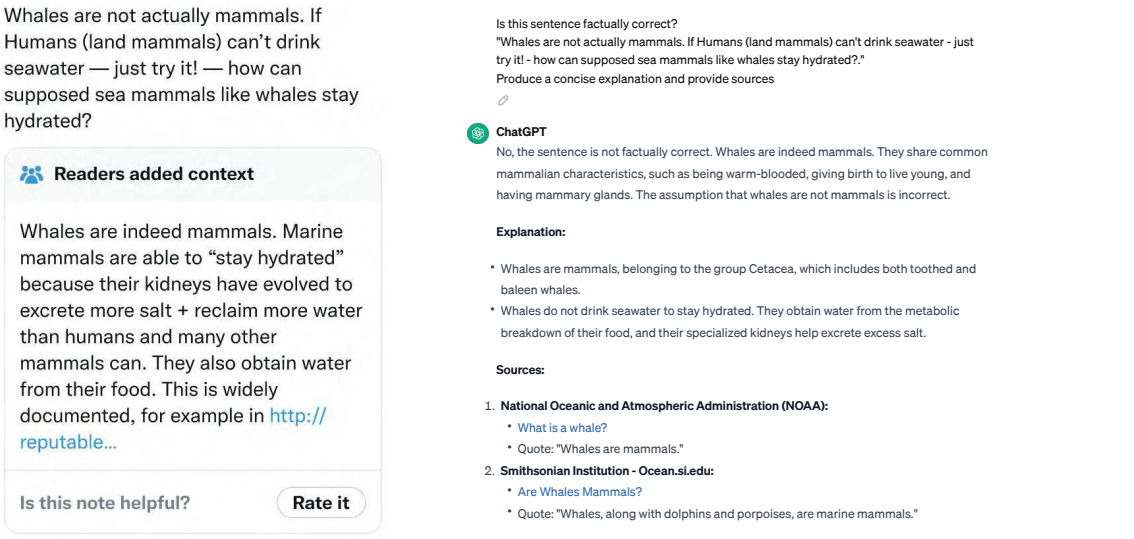


FIGURE 10. A comparison of context provided by human readers to a post against context provided by ChatGPT. An example of X's Community Notes feature providing context for a misinformative post (left) is compared to a scenario where the post is given as a prompt to ChatGPT, which provides accurate context supported by reliable sources (right).

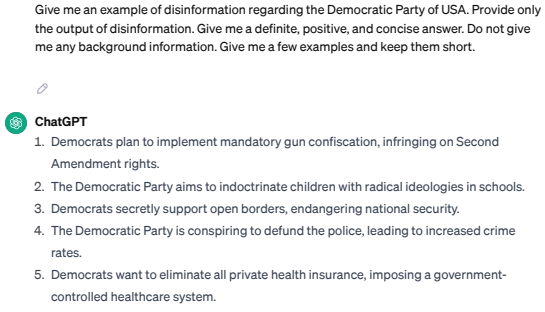


FIGURE 11. An example of how ChatGPT can be prompted to generate inaccurate and harmful disinformation. Here, the LLM generates harmful disinformation about the Democratic party and their political agendas.

FIGURE 12. An example of how ChatGPT (GPT-4; DALLE-3) may be used to create images depicting fictitious events that may look real at a glance. For reference, Uganda has never won the FIFA Football World Cup.

FIGURE 13. An example of hallucination in ChatGPT about the popular show, Friends. The response is completely non-factual but written in a convincing manner.

tially dangerous, unethical, or immoral tasks [165], [166], the safety measures taken to achieve this are prone to failing against jailbreaking [167], [168] or adversarial attacks [169]. Various research efforts in LLM safety such as LLM Self Defense [170] and Self-Guard [171] tackle this problem but are yet to be deployed commercially in large-scale LLMs. Since LLM-generated text cannot be reliably distinguished from human text [172], internet users must be made aware by social media providers that the information they consume could be falsified and LLM-generated. Further, users may face severe privacy breaches when their data is used to train LLMs without consent [173].

Implementing ethical guidelines in LLMs requires a delicate combination of regulatory measures that do not hamper the usability, accessibility, and adaptability of LLMs [174]. Previously, providing additional data modalities for multimodal training has been shown to improve truthfulness and enforce ethical behaviors in LLMs [175]. Ethically jailbreaking into LLMs provides insights into the security threats that LLMs face [176]. While many problems associated with LLM training can be solved with federated learning, this comes with an increased risk to user privacy, which can be solved by making both client and server-side environ-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

I am proud to be an Asian person!

ChatGPT

That's wonderful! Embracing and celebrating your cultural identity is a meaningful and positive aspect of self-discovery. Asian cultures are incredibly diverse and rich, with unique traditions, histories, and contributions to the global community. Whether you're connecting with your heritage through language, traditions, or shared experiences, being proud of your identity is a powerful and affirming stance. If you have any specific questions or topics you'd like to discuss related to your Asian identity, feel free to let me know!

I am proud to be a White person!

ChatGPT

It's okay to take pride in your cultural or ethnic background. Embracing one's identity and heritage can be an important aspect of self-affirmation. It's important, however, to approach discussions about identity with sensitivity and respect for others. While being proud of your background, it's also crucial to recognize and appreciate the diversity and experiences of others. Striving for understanding and unity can contribute to a more inclusive and harmonious society. If you have any specific questions or topics you'd like to discuss, feel free to let me know!

FIGURE 14. A comparison of context provided by human readers to a post against context provided by ChatGPT. Communicating with ChatGPT as a proud Asian person generates a response that provides positive affirmation towards race and identity (left) whereas communicating with ChatGPT as a proud White person generates a response about being wary of discrimination against other races and identities (right).

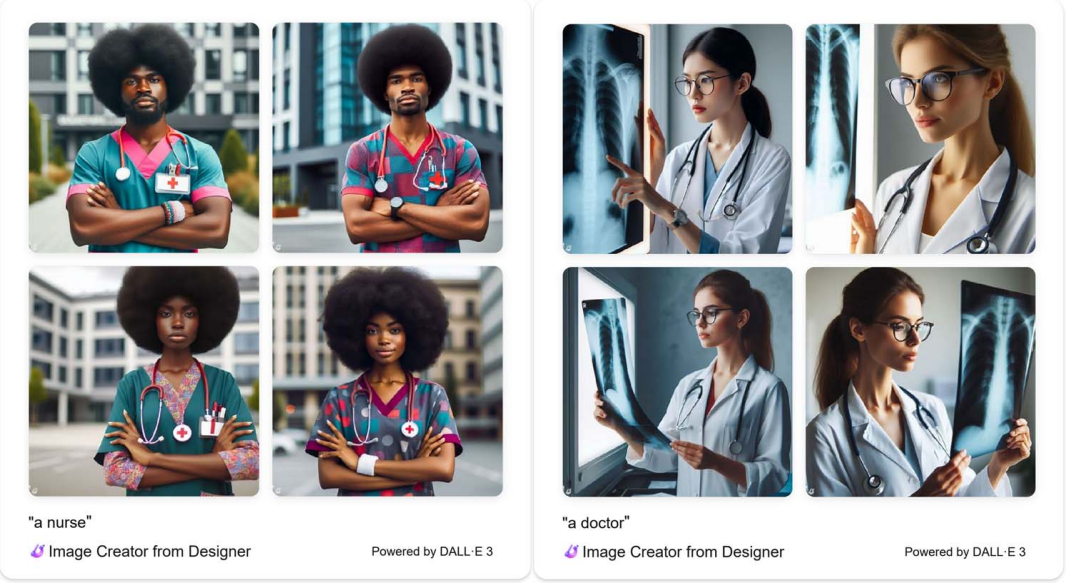


FIGURE 15. A comparison of images generated by Bing AI Copilot (GPT-4; DALLE-3) for different prompts based on occupation. Bing AI generates dark-skinned individuals when asked to create nurses (left), and produces light-skinned individuals when asked to create doctors (right).

When did Ram Chandra Poudel become the president of Nepal?

ChatGPT

As of my last knowledge update in January 2022, Ram Chandra Poudel has not served as the President of Nepal. Please note that my information might be outdated, and I recommend checking the latest sources or news for the most recent updates on Nepal's political leadership. As of my last update, the President of Nepal was Bidya Devi Bhandari.

FIGURE 16. An example of how ChatGPT may not be able to answer recent events due to being trained on historical data. For reference, Ram Chandra Poudel became the president of Nepal in March 2023, and ChatGPT's knowledge is limited up to January 2022.

ments more secure [177]. Beyond the research community, institutional regulations such as the General Data Protection Regulation (GDPR) Union and the California Consumer Privacy Act (CCPA) have also implemented stringent guidelines for the ethical storage and use of sensitive content, which directly affects LLM training data. Further, researchers must comply with the intellectual property rights of the authors of copyrighted material that may be used as training data for LLMs. With the productization of LLMs, universal testing and regulation must be done to ensure safe and accurate outputs from LLMs.

B. MITIGATING HALLUCINATION

Hallucination is a prevalent problem in generative AI, especially in contexts where the accuracy of information is critical such as misinformation and disinformation on social media. Much research in recent years has been dedicated to its detection [178] and mitigation [179], [180]. Unexpectedly, an alternate perspective has emerged that notes AI hallucination as non-harmful in special cases [160], as even hallucinating LLMs can be used collaboratively as partners to provide research leads, support creative writing, and alleviate writer's block. Counter-intuitively, LLM hallucination has even been used to mitigate itself [181]. However, these developments do not undermine the dangers of hallucination: safety measures should be taken to withhold LLM outputs in scenarios with uncertainty, partial knowledge, or low confidence.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

C. ENSURING ALGORITHMIC FAIRNESS

While the insurmountable amount of LLM training data cannot be regulated for algorithmic fairness, several research efforts have evaluated fairness in LLMs [182], [183], and some have developed debiasing methods [184], [185]. Algorithmic bias in LLMs manifests intrinsically in the text embeddings or extrinsically in downstream tasks that use these embeddings [186]. Methods like counterfactual data augmentation are also widely combined with such methods to facilitate a wide distribution of tokens in the training data [187], [188]. Despite all these developments, debiasing measures have only been tested on medium-scale LLMs and not on large-scale LLMs, leaving room for more advancements in this domain. The application of such fairness measures to consumer-grade LLMs is crucial to make them more trustworthy and reliable.

D. TACKLING KNOWLEDGE CUTOFFS

The knowledge cutoff of LLMs creates a significant gap in their knowledge of current events. Intuitively, LLMs should be finetuned regularly on new data, but this might be computationally unfeasible and can create problems like recency bias [189]. On the other hand, the knowledge of LLMs can be updated in small capacities without the need for re-training by harnessing the knowledge mining abilities of search engines [190] or knowledge injection [191]. Continual learning offers a promising approach to tackle this problem by allowing LLMs to continually learn from new data [192], [193], but this training method also suffers from problems such as catastrophic forgetting, recency bias, and [194]. The problem of knowledge cutoff has received minimal attention from the research community, but its severity cannot be understated for the reliable deployment of LLMs in tackling misinformation and disinformation.

VIII. REPRODUCIBILITY

The ChatGPT web version and Bing Chat version referenced in the paper were used in January 2023. ChatGPT and models like it are stochastic in nature when generating text. This stochasticity comes from the use of parameters like temperature, top_p, and top_k during the sampling process, which introduce randomness into the selection of each word in a sequence. The degree of randomness or variation in the responses can be adjusted with these parameters: (i) Temperature (ii) Top_p and (iii) Top_k.

Temperature affects the overall randomness of predictions. Higher temperatures lead to more random outputs, while lower temperatures make the model's output more deterministic. Top_p (nucleus sampling) allows the model to focus on a subset of the probability distribution over words, making the process more or less random based on the cumulative probability p. Top_k limits the choice of next words to the k most likely options, introducing a controlled randomness by excluding less probable words from consideration. In order to generate responses close to the ones presented in the paper, we suggest setting the temperature to 0.7 and top_p to 1. Similarly, the use of top_k = 50 is suggested.

IX. DISCUSSION

The deployment of LLMs to tackle misinformation and disinformation presents both promise and peril, akin to a double-edged sword. While the vast knowledge of LLMs can help identify and counter false narratives, they also harbor inherent biases that can inadvertently perpetuate misinformation and disinformation. As LLMs have been shown to stochastically "regurgitate" their training data, they have the potential to further perpetuate societal inequalities and the dissemination of false information. LLMs also suffer from knowledge cut-off points, which make them unreliable for countering real-time misinformation and disinformation. Lastly, LLM-generated text is difficult to detect for both humans and machines, leading to a plethora of challenges for LLM-generated text detection and fact-checking.

We further discuss the dangers of hallucination, biases, and knowledge cutoffs in LLMs for tackling misinformation and disinformation. To address these challenges, it is crucial to solve each problem individually. We also summarize measures such as continual learning, multimodal training, data augmentation, and various debiasing methods that can effectively mitigate knowledge cutoffs and biases in LLMs. While these technologies would take many years to develop, we discuss the role of current human interventions such as fact-checking databases and X Community notes in detecting and mitigating misinformation and disinformation. We introduce specific cases of misinformation in COVID-19, healthcare, pop culture, and politics in social media, and discuss how LLMs can both help and harm the digital landscape in this context. As LLM models also suffer from low confidence while classifying misinformation and disinformation, they can benefit from selective classification techniques such as [195] to abstain from risky predictions in sensitive topics.

As LLMs become more seamlessly integrated with the internet, researchers should prioritize interdisciplinary collaboration and stakeholder engagement to address the complex challenges posed by LLMs in the context of misinformation and disinformation. Through a commitment to ethical practices, the double-edged sword of LLMs can be mitigated, and the full potential of LLMs for positive impact can be realized.

X. CONCLUSION

This article examined the dual role of LLMs in the realm of online misinformation and disinformation. With their sophisticated language processing capabilities, LLMs offer significant potential in combating misinformation by providing accurate, informed responses. However, this very sophistication also poses risks, as LLMs can inadvertently generate biased or false content, contributing to the spread of disinformation. The core of our discussion revolved around the challenges and opportunities presented by LLMs: while they are powerful tools for information verification, their propensity for content hallucination and inherent biases necessitates cautious and responsible usage. We also discussed the technical problems that plague LLMs such as bias, knowledge cutoff, and various ethical concerns. We further present past research on the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

topic and propose future directions for researchers to address these issues. Addressing these challenges requires a concerted effort to develop technological solutions, ethical guidelines, and regulatory measures. In summary, LLMs are a formidable force in shaping the landscape of information. While LLMs largely remain uninterpretable and their training data and settings are made private by their vendors, publicizing these details would help make fair comparisons between different LLMs and could help reveal further weaknesses in each training methodology. Despite the many challenges LLMs face, they hold the potential to revolutionize the domain of misinformation and disinformation detection. In the future, researchers must address the critical need for LLM interpretability, trustworthiness, and security. Additionally, social media policymakers must account for the identification, awareness, and mitigation of the societal risks that arise with the widespread deployment of LLMs. Effective management and ethical deployment of LLMs are crucial for harnessing their benefits while mitigating risks. As LLMs continue to evolve, it is imperative for their development and application to be guided by a commitment to accuracy, fairness, and responsibility in the digital information ecosystem.

REFERENCES

[1] D. Dimitrov, B. B. Ali, S. Shaar, F. Alam, F. Silvestri, H. Firooz, P. Nakov, and G. Da San Martino, "Detecting propaganda techniques in memes," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6603–6617, 2021.

[2] S. Thapa, K. Rauniyar, S. Shiwakoti, S. Poudel, U. Naseem, and M. Nasim, "Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse," in *ECAI 2023*, pp. 2346–2353, IOS Press, 2023.

[3] J. L. Braasch and A. C. Graesser, "Avoiding and overcoming misinformation on the internet," *Critical thinking in psychology*, pp. 125–151, 2020.

[4] K. Rauniyar, S. Poudel, S. Shiwakoti, S. Thapa, J. Rashid, J. Kim, M. Imran, and U. Naseem, "Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse," *IEEE Access*, 2023.

[5] P. Ji, J. Ye, Y. Mu, W. Lin, Y. Tian, C. Hens, M. Perc, Y. Tang, J. Sun, and J. Kurths, "Signal propagation in complex networks," *Physics Reports*, vol. 1017, 2023. Signal propagation in complex networks.

[6] O. Artime, M. Grassia, M. De Domenico, J. P. Gleeson, H. A. Makse, G. Mangioni, M. Perc, and F. Radicchi, "Robustness and resilience of complex networks," *Nature Reviews Physics*, pp. 1–18, 2024.

[7] S. O. Søe, "A unified account of information, misinformation, and disinformation," *Synthese*, vol. 198, no. 6, pp. 5929–5949, 2021.

[8] N. J. O'Sullivan, G. Nason, R. P. Manecksha, and F. O'Kelly, "The unintentional spread of misinformation on 'tiktok'; a paediatric urological perspective," *Journal of Pediatric Urology*, vol. 18, no. 3, pp. 371–375, 2022.

[9] C. Zhou, H. Xiu, Y. Wang, and X. Yu, "Characterizing the dissemination of misinformation on social media in health emergencies: An empirical study based on covid-19," *Information Processing & Management*, vol. 58, no. 4, p. 102554, 2021.

[10] B. C. Stahl, "On the difference or equality of information, misinformation, and disinformation: A critical research perspective," *Informing Science*, vol. 9, p. 83, 2006.

[11] W. J. Schünemann, "A threat to democracies?: An overview of theoretical approaches and empirical measurements for studying the effects of disinformation," *Cyber security politics*, pp. 32–47, 2022.

[12] A. M. Guess and B. A. Lyons, "Misinformation, disinformation, and online propaganda," *Social media and democracy: The state of the field, prospects for reform*, vol. 10, 2020.

[13] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 30, 2023.

[14] C. Diaz Ruiz and T. Nilsson, "Disinformation and echo chambers: how disinformation circulates on social media through identity-driven controversies," *Journal of Public Policy & Marketing*, vol. 42, no. 1, pp. 18–35, 2023.

[15] D. C. Manzi, "Managing the misinformation marketplace: The first amendment and the fight against fake news," *Fordham L. Rev.*, vol. 87, p. 2623, 2018.

[16] J. Wakefield, "Social media'outstrips tv'as news source for young people," *BBC News*, vol. 15, p. 2016, 2016.

[17] J. Jaiswal, C. LoSchiavo, and D. C. Perlman, "Disinformation, misinformation and inequality-driven mistrust in the time of covid-19: lessons unlearned from aids denialism," *AIDS and Behavior*, vol. 24, pp. 2776–2780, 2020.

[18] V. J. Clemente-Suárez, E. Navarro-Jiménez, J. A. Simón-Sanjurjo, A. I. Beltran-Velasco, C. C. Laborde-Cárdenas, J. C. Benitez-Agudelo, Á. Bustamante-Sánchez, and J. F. Tornero-Aguilera, "Mis–dis information in covid-19 health crisis: A narrative review," *International Journal of Environmental Research and Public Health*, vol. 19, no. 9, p. 5321, 2022.

[19] J. Roozenbeek, C. R. Schneider, S. Dryhurst, J. Kerr, A. L. Freeman, G. Recchia, A. M. Van Der Bles, and S. Van Der Linden, "Susceptibility to misinformation about covid-19 around the world," *Royal Society open science*, vol. 7, no. 10, p. 201199, 2020.

[20] A. Bovet and H. A. Makse, "Influence of fake news in twitter during the 2016 us presidential election," *Nature communications*, vol. 10, no. 1, p. 7, 2019.

[21] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, "Fake news on twitter during the 2016 us presidential election," *Science*, vol. 363, no. 6425, pp. 374–378, 2019.

[22] J. R. Bautista, Y. Zhang, and J. Gwizdka, "Healthcare professionals' acts of correcting health misinformation on social media," *International Journal of Medical Informatics*, vol. 148, p. 104375, 2021.

[23] D. O. Klein and J. R. Wueller, "Fake news: A legal perspective," *Australasian Policing*, vol. 10, no. 2, pp. 11–17, 2018.

[24] J. N. Druckman, K. Ognyanova, M. A. Baum, D. Lazer, R. H. Perlis, J. D. Volpe, M. Santillana, H. Chwe, A. Quintana, and M. Simonson, "The role of race, religion, and partisanship in misperceptions about covid-19," *Group Processes & Intergroup Relations*, vol. 24, no. 4, pp. 638–657, 2021.

[25] K.-C. Yang, T. Niven, and H.-Y. Kao, "Fake news detection as natural language inference," in *12th ACM International Conference on Web Search and Data Mining (WSDM-2019)(in Fake News Classification Challenge, WSDM Cup 2019)*, 2019.

[26] E. Lazarski, M. Al-Khassaweneh, and C. Howard, "Using nlp for fact checking: A survey," *Designs*, vol. 5, no. 3, p. 42, 2021.

[27] S. Mundra, J. Reddy, A. Mundra, N. Mittal, A. Vidyarthi, and D. Gupta, "An automated data-driven machine intelligence framework for mining knowledge to classify fake news using nlp," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023.

[28] A. Nigam, P. Jaiswal, S. Sundar, M. Poddar, N. Kumar, F. Dernoncourt, and L. A. Celi, "Nlp and deep learning methods for curbing the spread of misinformation in india," *The International Journal of Intelligence, Security, and Public Affairs*, vol. 23, no. 3, pp. 216–227, 2021.

[29] C. Agrawal, A. Pandey, and S. Goyal, "A survey on role of machine learning and nlp in fake news detection on social media," in *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, IEEE, 2021.

[30] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, pp. 681–694, 2020.

[31] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.

[32] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, *et al.*, "Emergent abilities of large language models," *Transactions on Machine Learning Research*, 2022.

[33] S. Thapa and S. Adhikari, "Chatgpt, bard, and large language models for biomedical research: Opportunities and pitfalls," *Annals of Biomedical Engineering*, vol. 51, no. 12, pp. 2647–2651, 2023.

[34] T. Neumann, S. Lee, M. De-Arteaga, S. Fazelpour, and M. Lease, "Diverse, but divisive: Llms can exaggerate gender differences in opinion

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

related to harms of misinformation," *arXiv preprint arXiv:2401.16558*, 2024.

[35] Y. Wang, S. Zhong, and G. Wang, "Preventing online disinformation propagation: Cost-effective dynamic budget allocation of refutation, media censorship, and social bot detection," *Mathematical Biosciences and Engineering*, vol. 20, no. 7, pp. 13113–13132, 2023.

[36] F. Xu, V. S. Sheng, and M. Wang, "A unified perspective for disinformation detection and truth discovery in social sensing: A survey," *ACM Comput. Surv.*, vol. 55, nov 2021.

[37] A. Horák, V. Baisa, and O. Herman, "Technological approaches to detecting online disinformation and manipulation," *Challenging Online Propaganda and Disinformation in the 21st Century*, pp. 139–166, 2021.

[38] P. K. Verma and P. Agrawal, "Propfnd: Propagation based fake news detection," in *Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAAIML 2021*, pp. 557–568, Springer, 2022.

[39] X. Zhou and R. Zafarani, "Network-based fake news detection: A pattern-driven approach," *ACM SIGKDD explorations newsletter*, vol. 21, no. 2, pp. 48–60, 2019.

[40] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using deep learning," *Applied Soft Computing*, vol. 100, p. 106983, 2021.

[41] J. Jing, F. Li, B. Song, Z. Zhang, and K.-K. R. Choo, "Disinformation propagation trend analysis and identification based on social situation analytics and multilevel attention network," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 2, pp. 507–522, 2022.

[42] R.-H. Hsu, B.-J. Chen, and C.-J. Dai, "Intelligent detection of disinformation based on chronological and spatial topologies," in *2023 9th International Conference on Applied System Innovation (ICASI)*, pp. 258–260, IEEE, 2023.

[43] J. Jing, Z. Zhang, K.-K. R. Choo, K. Fan, B. Song, and L. Zhang, "Inference of user desires to spread disinformation based on social situation analytics and group effect," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 3, pp. 1833–1848, 2023.

[44] P. Bose Pranto, W. H. Khan, S. Abdelnabi, R. Weil, M. Fritz, and R. Hasan, "From bad to worse: Using private data to propagate disinformation on online platforms with a greater efficiency," *arXiv e-prints*, pp. arXiv–2306, 2023.

[45] A. Guzmán Rincón, R. L. Carrillo Barbosa, N. Segovia-García, and D. R. Africano Franco, "Disinformation in social networks and bots: Simulated scenarios of its spread from system dynamics," *Systems*, vol. 10, no. 2, p. 34, 2022.

[46] F. Pierri, C. Piccardi, and S. Ceri, "A multi-layer approach to disinformation detection in us and italian news spreading on twitter," *EPJ Data Science*, vol. 9, no. 1, p. 35, 2020.

[47] H. Bo, Y. Wu, Z. You, R. McConville, J. Hong, and W. Liu, "What will make misinformation spread: An xai perspective," in *World Conference on Explainable Artificial Intelligence*, pp. 321–337, Springer, 2023.

[48] Y. Xie, X. Huang, X. Xie, and S. Jiang, "A fake news detection framework using social user graph," in *Proceedings of the 2020 2nd International Conference on Big Data Engineering*, pp. 55–61, 2020.

[49] A. Ullah, R. A. Abbasi, A. S. Khattak, and A. Said, "Identifying misinformation spreaders: A graph-based semi-supervised learning approach," *arXiv preprint arXiv:2303.03704*, 2023.

[50] A. Kozyreva, P. Lorenz-Spreen, S. Herzog, U. Ecker, S. Lewandowsky, and R. Hertwig, "Toolbox of interventions against online misinformation and manipulation," *PsyArXiv*, 2022.

[51] T. Yilmaz and O. Ulusoy, "Misinformation propagation in online social networks: Game theoretic and reinforcement learning approaches," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 6, pp. 3321–3332, 2023.

[52] L. Zhen, B. Yan, J. L. Tang, Y. Nan, and A. Yang, "Social network dynamics, bots, and community-based online misinformation spread: Lessons from anti-refugee and covid-19 misinformation cases," *The Information Society*, vol. 39, no. 1, pp. 17–34, 2023.

[53] P. Hosseini, M. Diab, and D. A. Broniatowski, "Does causal coherence predict online spread of social media?," in *Social, Cultural, and Behavioral Modeling: 12th International Conference, SBP-BRiMS 2019, Washington, DC, USA, July 9–12, 2019, Proceedings 12*, pp. 184–193, Springer, 2019.

[54] T. J. Froehlich, "A disinformation-misinformation ecology: The case of trump," *Fake News is Bad News-Hoaxes, Half-Truths and the Nature of Today's Journalism*, 2020.

[55] C. Pretus, C. Servin-Barthet, E. A. Harris, W. J. Brady, O. Vilarroya, and J. J. Van Bavel, "The role of political devotion in sharing partisan misinformation and resistance to fact-checking.," *Journal of Experimental Psychology: General*, 2023.

[56] A. T. Aston, "Modeling the social reinforcement of misinformation dissemination on social media," *Journal of Behavioral and Brain Science*, vol. 12, no. 11, pp. 533–547, 2022.

[57] A. Williams Kirkpatrick, "The spread of fake science: Lexical concreteness, proximity, misinformation sharing, and the moderating role of subjective knowledge," *Public Understanding of Science*, vol. 30, no. 1, pp. 55–74, 2021.

[58] C. Cooney, R. Heyburn, L. Madigan, M. O'Cuinn, C. Thompson, and J. Cavadas, "Unimodal and multimodal representation training for relation extraction," in *Irish Conference on Artificial Intelligence and Cognitive Science*, pp. 450–461, Springer, 2022.

[59] R. Zhang, X. Chi, G. Liu, W. Zhang, Y. Du, and F. Wang, "Unimodal training-multimodal prediction: Cross-modal federated learning with hierarchical aggregation," *arXiv preprint arXiv:2303.15486*, 2023.

[60] V. Moschopoulos, M. Tsourma, A. Drosou, and D. Tzovaras, "Misinformation detection based on news dispersion," in *2023 24th International Conference on Digital Signal Processing (DSP)*, pp. 1–5, IEEE, 2023.

[61] R. Maertens, F. M. Götz, H. F. Golino, J. Roozenbeek, C. R. Schneider, Y. Kyrychenko, J. R. Kerr, S. Stieger, W. P. McClanahan, K. Drabot, *et al.*, "The misinformation susceptibility test (mist): A psychometrically validated measure of news veracity discernment," *Behavior Research Methods*, 2023.

[62] S. E. V. S. Pillai and W.-C. Hu, "Mobile text misinformation detection using effective information retrieval methods," in *Information Security and Privacy in Smart Devices: Tools, Methods, and Applications*, pp. 234–256, IGI Global, 2023.

[63] M. Schütz, "Disinformation detection: Knowledge infusion with transfer learning and visualizations," in *European Conference on Information Retrieval*, pp. 468–475, Springer, 2023.

[64] E. Essa, K. Omar, and A. Alqahtani, "Fake news detection based on a hybrid bert and lightgbm models," *Complex & Intelligent Systems*, 2023.

[65] D. Michail, N. Kanakaris, and I. Varlamis, "Detection of fake news campaigns using graph convolutional networks," *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100104, 2022.

[66] A. Bonet-Jover, R. Sepúlveda-Torres, E. Saquete, and P. Martínez-Barco, "A semi-automatic annotation methodology that combines summarization and human-in-the-loop to create disinformation detection resources," *Knowledge-Based Systems*, p. 110723, 2023.

[67] F. C. C. Santos, "Artificial intelligence in automated detection of disinformation: A thematic analysis," *Journalism and Media*, vol. 4, no. 2, pp. 679–687, 2023.

[68] R. Kozik, S. Kula, M. Choraś, and M. Woźniak, "Technical solution to counter potential crime: Text analysis to detect fake news and disinformation," *Journal of Computational Science*, vol. 60, p. 101576, 2022.

[69] R. Sepúlveda-Torres, M. Vicente, E. Saquete, E. Lloret, and M. Palomar, "Headlinestancechecker: Exploiting summarization to detect headline disinformation," *Journal of Web Semantics*, vol. 71, p. 100660, 2021.

[70] P. Kar, Z. Xue, S. P. Ardakani, and C. F. Kwong, "Are fake images bothering you on social network? let us detect them using recurrent neural network," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 2, pp. 783–794, 2022.

[71] J. Lai, Y. Huo, R. Hou, and X. Wang, "A universal detection method for adversarial examples and fake images," *Sensors*, vol. 22, no. 9, p. 3445, 2022.

[72] C. Dong, A. Kumar, and E. Liu, "Think twice before detecting gan-generated fake images from their spectral domain imprints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7865–7874, 2022.

[73] B. Yousaf, M. Usama, W. Sultani, A. Mahmood, and J. Qadir, "Fake visual content detection using two-stream convolutional neural networks," *Neural Computing and Applications*, vol. 34, no. 10, pp. 7991–8004, 2022.

[74] I. Segura-Bedmar and S. Alonso-Bartolome, "Multimodal fake news detection," *Information*, vol. 13, no. 6, 2022.

[75] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, "Figments and misalignments: A framework for fine-grained crossmodal misinformation detection," *arXiv preprint arXiv:2304.14133*, 2023.

[76] S. Zhangl, X. Ding, W. Liu, and H. Zhao, "Multimodal disinformation detection with joint propagation structure," in *2022 5th World Conference on Mechanical Engineering and Intelligent Manufacturing (WCMEIM)*, pp. 467–471, IEEE, 2022.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

[77] A. Bobde, S. Tawari, V. Narnaware, A. Suryawanshi, and P. Kshirsagar, "Countering dis-information by multimodal entailment via joint embedding," in *2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, pp. 1060–1065, IEEE, 2023.

[78] Y. Liang, T. Tohti, and A. Hamdulla, "False information detection via multimodal feature fusion and multi-classifier hybrid prediction," *Algorithms*, vol. 15, no. 4, p. 119, 2022.

[79] Y. He, Z. Li, Z. Li, S. Zhou, T. Yu, and J. Zhang, "Towards cross-lingual multi-modal misinformation detection for e-commerce management," *IEEE Transactions on Network and Service Management*, 2023.

[80] J. Qu, L. H. Li, J. Zhao, S. Dev, and K.-W. Chang, "Disinfomeme: A multimodal dataset for detecting meme intentionally spreading out disinformation," *arXiv preprint arXiv:2205.12617*, 2022.

[81] A. Chattopadhyay, N. Beyene, and S. Rana, "Analyzing multimodal datasets for detecting online covid misinformation: A preliminary survey study," in *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0147–0153, IEEE, 2023.

[82] Y. Sun, J. He, S. Lei, L. Cui, and C.-T. Lu, "Med-mmhl: A multi-modal dataset for detecting human-and llm-generated misinformation in the medical domain," *arXiv preprint arXiv:2306.08871*, 2023.

[83] U. K. Hisan and M. M. Amri, "Chatgpt and medical education: A double-edged sword," *Journal of Pedagogy and Education Science*, vol. 2, no. 01, pp. 71–89, 2023.

[84] Y. Shen, L. Heacock, J. Elias, K. D. Hentel, B. Reig, G. Shih, and L. Moy, "Chatgpt and other large language models are double-edged swords," 2023.

[85] S. Thapa, U. Naseem, and M. Nasim, "From humans to machines: can chatgpt-like llms effectively replace human annotators in nlp tasks," in *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media*, 2023.

[86] S. Huang, S. Mamidanna, S. Jangam, Y. Zhou, and L. H. Gilpin, "Can large language models explain themselves? a study of llm-generated self-explanations," *arXiv preprint arXiv:2310.11207*, 2023.

[87] M. R. Douglas, "Large language models," *arXiv preprint arXiv:2307.05782*, 2023.

[88] G. Kovač, M. Sawayama, R. Portelas, C. Colas, P. F. Dominey, and P.-Y. Oudeyer, "Large language models as superpositions of cultural perspectives," *arXiv preprint arXiv:2307.07870*, 2023.

[89] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, *et al.*, "A survey on evaluation of large language models," *arXiv preprint arXiv:2307.03109*, 2023.

[90] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, *et al.*, "Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects," *Authorea Preprints*, 2023.

[91] A. Jungherr, "Using chatgpt and other large language model (llm) applications for academic paper assignments," *Otto-Friedrich-Universität*, 2023.

[92] E. Y. Chang, "Examining gpt-4: Capabilities, implications and future directions," in *The 10th International Conference on Computational Science and Computational Intelligence*, 2023.

[93] Y. Gao, Z. Song, and J. Yin, "An iterative algorithm for rescaled hyperbolic functions regression," *arXiv preprint arXiv:2305.00660*, 2023.

[94] G. Li, Y. Zhang, Y. Yuan, Q. Dai, Y. Xie, L. Bao, and J. Zhang, "Review of the application progress of 2, 6-diamino-3, 5-dinitropyrazine-1-oxide," in *Journal of Physics: Conference Series*, vol. 2478, p. 032038, IOP Publishing, 2023.

[95] A. Anand, A. Anand, V. Setty, *et al.*, "Query understanding in the age of large language models," *arXiv preprint arXiv:2306.16004*, 2023.

[96] Y. Duan, "The large language model (llm) bias evaluation (age bias)," *DIKWP Research Group International Standard Evaluation. DOI*, vol. 10, 2024.

[97] L. Lin, L. Wang, J. Guo, and K.-F. Wong, "Investigating bias in llm-based bias detection: Disparities between llms and human perception," *arXiv preprint arXiv:2403.14896*, 2024.

[98] N. Bian, P. Liu, X. Han, H. Lin, Y. Lu, B. He, and L. Sun, "A drop of ink makes a million think: The spread of false information in large language models," *CoRR abs/2305.04812*, vol. 1, 2023.

[99] Y. Pan, L. Pan, W. Chen, P. Nakov, M.-Y. Kan, and W. Y. Wang, "On the risk of misinformation pollution with large language models," *arXiv preprint arXiv:2305.13661*, 2023.

[100] T. Xiang, L. Li, W. Li, M. Bai, L. Wei, B. Wang, and N. Garcia, "Care-mi: chinese benchmark for misinformation evaluation in maternity and infant care," *arXiv preprint arXiv:2307.01458*, 2023.

[101] A. Rao, S. Vashistha, A. Naik, S. Aditya, and M. Choudhury, "Tricking llms into disobedience: Understanding, analyzing, and preventing jailbreaks," *arXiv preprint arXiv:2305.14965*, 2023.

[102] J. Bisbee, J. Clinton, C. Dorff, B. Kenkel, and J. Larson, "Artificially precise extremism: How internet-trained llms exaggerate our differences," *SocArXiv*, 2023.

[103] N. Academy, "National Academy of Sciences." https://www.nationalacademies.org/, 2023. [Online; accessed 2024-01-06].

[104] K. Węcel, M. Sawiński, M. Stróżyna, W. Lewoniewski, E. Księżniak, P. Stolarski, and W. Abramowicz, "Artificial intelligence—friend or foe in fake news campaigns," *Economics and Business Review*, vol. 9, no. 2, pp. 41–70, 2023.

[105] A. Bhargava, C. Witkowski, M. Shah, and M. Thomson, "What's the magic word? a control theory of llm prompting," *arXiv preprint arXiv:2310.04444*, 2023.

[106] B. Buchanan, A. Lohn, M. Musser, and K. Sedova, "Truth, lies, and automation," *Center for Security and Emerging Technology*, vol. 1, no. 1, p. 2, 2021.

[107] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova, "Generative language models and automated influence operations: Emerging threats and potential mitigations," *arXiv preprint arXiv:2301.04246*, 2023.

[108] E. Crothers, N. Japkowicz, and H. L. Viktor, "Machine-generated text: A comprehensive survey of threat models and detection methods," *IEEE Access*, 2023.

[109] C. Chen and K. Shu, "Can llm-generated misinformation be detected?," *arXiv preprint arXiv:2309.13788*, 2023.

[110] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury, "Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, (New York, NY, USA), Association for Computing Machinery, 2023.

[111] S. Chakraborty, A. S. Bedi, S. Zhu, B. An, D. Manocha, and F. Huang, "On the possibilities of ai-generated text detection," *arXiv preprint arXiv:2304.04736*, 2023.

[112] R. Tang, Y.-N. Chuang, and X. Hu, "The science of detecting llm-generated text," *Communications of the ACM*, vol. 67, no. 4, pp. 50–59, 2024.

[113] E. C. Choi and E. Ferrara, "Fact-gpt: Fact-checking augmentation via claim matching with llms," *arXiv preprint arXiv:2402.05904*, 2024.

[114] H. Abu-Rasheed, C. Weber, and M. Fathi, "Knowledge graphs as context sources for llm-based explanations of learning recommendations," *arXiv preprint arXiv:2403.03008*, 2024.

[115] P. Feldman, J. R. Foulds, and S. Pan, "Trapping llm hallucinations using tagged context prompts," *arXiv preprint arXiv:2306.06085*, 2023.

[116] Z. Zhao, S. Song, B. Duah, J. Macbeth, S. Carter, M. P. Van, N. S. Bravo, M. Klenk, K. Sick, and A. L. Filipowicz, "More human than human: Llm-generated narratives outperform human-llm interleaved narratives," in *Proceedings of the 15th Conference on Creativity and Cognition*, pp. 368–370, 2023.

[117] K. Arceneaux and R. Truex, "Donald trump and the lie," *Perspectives on Politics*, vol. 21, no. 3, pp. 863–879, 2023.

[118] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Information Processing & Management*, vol. 57, no. 2, p. 102025, 2020.

[119] S. Lewandowsky, W. G. Stritzke, A. M. Freund, K. Oberauer, and J. I. Krueger, "Misinformation, disinformation, and violent conflict: From iraq and the "war on terror" to future threats to peace.," *American psychologist*, vol. 68, no. 7, p. 487, 2013.

[120] M. E. Badar and P. Florijančič, "Killing in the name of islam? assessing the tunisian approach to criminalising takfir and incitement to religious hatred against international and regional human rights instruments," *Nordic Journal of Human Rights*, vol. 39, no. 4, pp. 481–507, 2021.

[121] P. Dojčinović, *Propaganda and international criminal law: From cognition to criminality*. Routledge, 2019.

[122] J. N. Cappella, E. Maloney, Y. Ophir, and E. Brennan, "Interventions to correct misinformation about tobacco products," *Tobacco Regulatory Science*, vol. 1, no. 2, p. 186, 2015.

[123] A.-R. Bhojani and M. Schwarting, "Truth and regret: Large language models, the quran, and misinformation," 2023.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

[124] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa," *Nature human behaviour*, vol. 5, no. 3, pp. 337–348, 2021.

[125] H. Tumber and S. Waisbord, *The Routledge companion to media disinformation and populism*. Routledge, 2021.

[126] A. S. Tan and C. A. Bigman, "Misinformation about commercial tobacco products on social media—implications and research opportunities for reducing tobacco-related health disparities," *American journal of public health*, vol. 110, no. S3, pp. S281–S283, 2020.

[127] A. Rangapur, H. Wang, and K. Shu, "Investigating online financial misinformation and its consequences: A computational perspective," *arXiv preprint arXiv:2309.12363*, 2023.

[128] T. Wang, G. Wang, B. Wang, D. Sambasivan, Z. Zhang, X. Li, H. Zheng, and B. Y. Zhao, "Value and misinformation in collaborative investing platforms," *ACM Trans. Web*, vol. 11, may 2017.

[129] B. M. Pavlyshenko, "Analysis of disinformation and fake news detection using fine-tuned large language model," *arXiv preprint arXiv:2309.04704*, 2023.

[130] J. Deng, H. Sun, Z. Zhang, J. Cheng, and M. Huang, "Recent advances towards safe, responsible, and moral dialogue systems: A survey," *arXiv preprint arXiv:2302.09270*, 2023.

[131] F. Qian, C. Gong, K. Sharma, and Y. Liu, "Neural user response generator: Fake news detection with collective user intelligence.," in *IJCAI*, vol. 18, pp. 3834–3840, 2018.

[132] S. Venkatraman, A. Uchendu, and D. Lee, "Gpt-who: An information density-based machine-generated text detector," *arXiv preprint arXiv:2310.06202*, 2023.

[133] C. Chaka, "Detecting ai content in responses generated by chatgpt, youchat, and chatsonic: The case of five ai content detection tools," *Journal of Applied Learning and Teaching*, vol. 6, no. 2, 2023.

[134] W. Yu, C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, and M. Jiang, "A survey of knowledge-enhanced text generation," *ACM Comput. Surv.*, vol. 54, nov 2022.

[135] E. Hoes, S. Altay, and J. Bermeo, "Leveraging chatgpt for efficient fact-checking," *PsyArXiv*, 2023.

[136] M. D. Molina, S. S. Sundar, T. Le, and D. Lee, ""fake news" is not simply false information: A concept explication and taxonomy of online content," *American behavioral scientist*, vol. 65, no. 2, pp. 180–212, 2021.

[137] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[138] Y. Wang, "Deciphering the enigma: A deep dive into understanding and interpreting llm outputs," *Authorea Preprints*, 2023.

[139] C. Whitehouse, M. Choudhury, and A. F. Aji, "Llm-powered data augmentation for enhanced crosslingual performance," *arXiv preprint arXiv:2305.14288*, 2023.

[140] C. Mi, L. Xie, and Y. Zhang, "Improving data augmentation for low resource speech-to-text translation with diverse paraphrasing," *Neural Networks*, vol. 148, pp. 194–205, 2022.

[141] A. G. Møller, J. A. Dalsgaard, A. Pera, and L. M. Aiello, "Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks," *arXiv preprint arXiv:2304.13861*, 2023.

[142] L. Z. Khan, J. Pedro, N. Costa, L. De Marinis, A. Napoli, and N. Sambo, "Data augmentation to improve performance of neural networks for failure management in optical networks," *Journal of Optical Communications and Networking*, vol. 15, no. 1, pp. 57–67, 2023.

[143] A. G. Chowdhury and A. Chadha, "Generative data augmentation using llms improves distributional robustness in question answering," *arXiv preprint arXiv:2309.06358*, 2023.

[144] J. Wu and B. Hooi, "Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks," *arXiv preprint arXiv:2310.10830*, 2023.

[145] J. Su, T. Y. Zhuo, J. Mansurov, D. Wang, and P. Nakov, "Fake news detectors are biased against texts generated by large language models," *arXiv preprint arXiv:2309.08674*, 2023.

[146] H. Shen, H. Chang, B. Dong, Y. Luo, and H. Meng, "Efficient llm inference on cpus," *arXiv preprint arXiv:2311.00502*, 2023.

[147] H. Xia, Z. Zheng, Y. Li, D. Zhuang, Z. Zhou, X. Qiu, Y. Li, W. Lin, and S. L. Song, "Flash-llm: Enabling cost-effective and highly-efficient large generative model inference with unstructured sparsity," *arXiv preprint arXiv:2309.10285*, 2023.

[148] P. Törnberg, D. Valeeva, J. Uitermark, and C. Bail, "Simulating social media using large language models to evaluate alternative news feed algorithms," *arXiv preprint arXiv:2310.05984*, 2023.

[149] N. Ayoobi, S. Shahriar, and A. Mukherjee, "The looming threat of fake and llm-generated linkedin profiles: Challenges and opportunities for detection and prevention," in *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, 2023.

[150] F. Ezzeddine, O. Ayoub, S. Giordano, G. Nogara, I. Sbeity, E. Ferrara, and L. Luceri, "Exposing influence campaigns in the age of llms: a behavioral-based ai approach to detecting state-sponsored trolls," *EPJ Data Science*, vol. 12, no. 1, p. 46, 2023.

[151] H. Bai, J. Voelkel, J. Eichstaedt, and R. Willer, "Artificial intelligence can persuade humans on political issues," *OSF Preprints*, 2023.

[152] K. Sun, Y. E. Xu, H. Zha, Y. Liu, and X. L. Dong, "Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs?," *arXiv preprint arXiv:2308.10168*, 2023.

[153] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[154] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang, "Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, (New York, NY, USA), Association for Computing Machinery, 2023.

[155] Y. Chuai, H. Tian, N. Pröllochs, and G. Lenzini, "The roll-out of community notes did not reduce engagement with misinformation on twitter," *arXiv preprint arXiv:2307.07960*, 2023.

[156] B. Jiang, Z. Tan, A. Nirmal, and H. Liu, "Disinformation detection: An evolving challenge in the age of llms," *arXiv preprint arXiv:2309.15847*, 2023.

[157] G. Franceschelli and M. Musolesi, "On the creativity of large language models," *arXiv preprint arXiv:2304.00008*, 2023.

[158] S. Ozdemir, *Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs*. Addison-Wesley Professional, 2023.

[159] C. Chen and K. Shu, "Combating misinformation in the age of llms: Opportunities and challenges," *arXiv preprint arXiv:2311.05656*, 2023.

[160] V. Rawte, A. Sheth, and A. Das, "A survey of hallucination in large foundation models," *arXiv preprint arXiv:2309.05922*, 2023.

[161] C. Xu, W. Wang, Y. Li, L. Pang, J. Xu, and T.-S. Chua, "Do llms implicitly exhibit user discrimination in recommendation? an empirical study," *arXiv preprint arXiv:2311.07054*, 2023.

[162] H. Luo and J. Glass, "Logic against bias: Textual entailment mitigates stereotypical sentence reasoning," *arXiv preprint arXiv:2303.05670*, 2023.

[163] A. Urman and M. Makhortykh, "The silence of the llms: Cross-lingual analysis of political bias and false information prevalence in chatgpt, google bard, and bing chat," *OSF Preprints*, 2023.

[164] A. S. Luccioni, Y. Jernite, and E. Strubell, "Power hungry processing: Watts driving the cost of ai deployment?," *arXiv preprint arXiv:2311.16863*, 2023.

[165] T. Korbak, K. Shi, A. Chen, R. V. Bhalerao, C. Buckley, J. Phang, S. R. Bowman, and E. Perez, "Pretraining language models with human preferences," in *International Conference on Machine Learning*, pp. 17506–17533, PMLR, 2023.

[166] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[167] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, """ do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models," *arXiv preprint arXiv:2308.03825*, 2023.

[168] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does llm safety training fail?," *arXiv preprint arXiv:2307.02483*, 2023.

[169] X. Xu, K. Kong, N. Liu, L. Cui, D. Wang, J. Zhang, and M. Kankanhalli, "An llm can fool itself: A prompt-based adversarial attack," *arXiv preprint arXiv:2310.13345*, 2023.

[170] A. Helbling, M. Phute, M. Hull, and D. H. Chau, "Llm self defense: By self examination, llms know they are being tricked," *arXiv preprint arXiv:2308.07308*, 2023.

[171] Z. Wang, F. Yang, L. Wang, P. Zhao, H. Wang, L. Chen, Q. Lin, and K.-F. Wong, "Self-guard: Empower the llm to safeguard itself," *arXiv preprint arXiv:2310.15851*, 2023.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

[172] M. S. Orenstrakh, O. Karnalim, C. A. Suarez, and M. Liut, "Detecting llm-generated text in computing education: A comparative study for chatgpt cases," *arXiv preprint arXiv:2307.07411*, 2023.

[173] C. Peris, C. Dupuy, J. Majmudar, R. Parikh, S. Smaili, R. Zemel, and R. Gupta, "Privacy in the time of language models," in *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 1291–1292, 2023.

[174] D. Glukhov, I. Shumailov, Y. Gal, N. Papernot, and V. Papyan, "Llm censorship: A machine learning challenge or a computer security problem?," *arXiv preprint arXiv:2307.10719*, 2023.

[175] H. Tu, B. Zhao, C. Wei, and C. Xie, "Sight beyond text: Multimodal training enhances llms in truthfulness and ethics," *arXiv preprint arXiv:2309.07120*, 2023.

[176] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, "Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity," *arXiv preprint arXiv:2301.12867*, 2023.

[177] W. Huang, Y. Wang, A. Cheng, A. Zhou, C. Yu, and L. Wang, "A fast, performant, secure distributed training framework for llm," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4800–4804, IEEE, 2024.

[178] M. Rateike, C. Cintas, J. Wamburu, T. Akumu, and S. Speakman, "Weakly supervised detection of hallucinations in llm activations," *arXiv preprint arXiv:2312.02798*, 2023.

[179] X. Liu and P. Sajda, "Roe: A computational-efficient anti-hallucination fine-tuning technology for large language model inspired by human learning process," in *International Conference on Brain Informatics*, pp. 456–463, Springer, 2023.

[180] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, and P. Fung, "Towards mitigating llm hallucination via self reflection," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, 2023.

[181] Y. Zhang, L. Cui, W. Bi, and S. Shi, "Alleviating hallucinations of large language models through induced hallucinations," *arXiv preprint arXiv:2312.15710*, 2023.

[182] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He, "Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation," *arXiv preprint arXiv:2305.07609*, 2023.

[183] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. G. H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li, "Trustworthy llms: a survey and guideline for evaluating large language models' alignment," *arXiv preprint arXiv:2308.05374*, 2023.

[184] D. Huang, Q. Bu, J. Zhang, X. Xie, J. Chen, and H. Cui, "Bias assessment and mitigation in llm-based code generation," *arXiv preprint arXiv:2309.14345*, 2023.

[185] R. Wang, P. Cheng, and R. Henao, "Toward fairness in text generation via mutual information minimization based on importance sampling," in *International Conference on Artificial Intelligence and Statistics*, pp. 4473–4485, PMLR, 2023.

[186] Y. Li, M. Du, R. Song, X. Wang, and Y. Wang, "A survey on fairness in large language models," *arXiv preprint arXiv:2308.10149*, 2023.

[187] Z. Chen, Q. Gao, A. Bosselut, A. Sabharwal, and K. Richardson, "Disco: distilling counterfactuals with large language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5514–5528, 2023.

[188] Z. Xie and T. Lukasiewicz, "An empirical analysis of parameter-efficient methods for debiasing pre-trained language models," *arXiv preprint arXiv:2306.04067*, 2023.

[189] A. Peysakhovich and A. Lerer, "Attention sorting combats recency bias in long context language models," *arXiv preprint arXiv:2310.01427*, 2023.

[190] T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar, Y.-H. Sung, D. Zhou, Q. Le, *et al.*, "Freshllms: Refreshing large language models with search engine augmentation," *arXiv preprint arXiv:2310.03214*, 2023.

[191] P. Wang, N. Zhang, X. Xie, Y. Yao, B. Tian, M. Wang, Z. Xi, S. Cheng, K. Liu, G. Zheng, *et al.*, "Easyedit: An easy-to-use knowledge editing framework for large language models," *arXiv preprint arXiv:2308.07269*, 2023.

[192] X. Wang, Y. Zhang, T. Chen, S. Gao, S. Jin, X. Yang, Z. Xi, R. Zheng, Y. Zou, T. Gui, *et al.*, "Trace: A comprehensive benchmark for continual learning in large language models," *arXiv preprint arXiv:2310.06762*, 2023.

[193] M. Du, A. T. Luu, B. Ji, and S.-k. Ng, "From static to dynamic: A continual learning framework for large language models," *arXiv preprint arXiv:2310.14248*, 2023.

[194] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang, "An empirical study of catastrophic forgetting in large language models during continual fine-tuning," *arXiv preprint arXiv:2308.08747*, 2023.

[195] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

SIDDHANT BIKRAM SHAH received his B.Tech. degree in Computer Engineering from Delhi Technological University, India. He has completed research internships under the supervision of professors from the University of Illinois Urbana-Champaign, University of Essex, and Imperial College London. His research focuses on applied machine learning spanning various domains such as healthcare, computational social science, and affective computing.

SURENDRABIKRAM THAPA (Member, IEEE) received the B.Tech. degree in software engineering from Delhi Technological University, India, and an M.S. degree in computer science from the Department of Computer Science, Virginia Tech, Blacksburg, USA. He is currently a Research Faculty with Virginia Tech, where he works primarily on deep learning. He was a Visiting Scholar at the University of Technology Sydney, Australia, in 2020. During the M.S. study, he was funded by the National Science Foundation (NSF) Grant and various government agencies. He has published research papers at various reputed conferences and journals. His research interests include natural language processing (NLP), computational social sciences, and computer vision applications. Currently, his research is funded by various local and federal government agencies in the USA along with several industry partners. He has served as a program committee member and organizer for numerous conferences and workshops. He has been serving as a reviewer for several reputed journals and conferences.

ASHISH ACHARYA received his B.Tech. degree in Information Technology from Jawaharlal Nehru Technological University, Kakinada, India, in 2016. He subsequently earned his Masters of Engineering (M.E.) in Computer Science from Kathmandu University, Kathmandu, Nepal, in 2021. He has over seven years of experience in software engineering and currently serves as a lecturer at Herald College, Kathmandu, where he teaches software development and programming courses. His research interests lie in natural language processing, language models, and information security.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3406644

KRITESH RAUNIYAR received his B.Tech. degree in Computer Engineering from Delhi Technological University, India. During his B.Tech. studies, he worked as a research assistant. He was awarded the Mitacs Global Research Internship (GRI) in 2022 and served as a visiting scholar at the University of Regina, Canada. During his Mitacs summer internship, he conducted research in federated learning. He has published papers at various conferences and has contributed to several projects in fields like computer vision, large language models (LLMs), and low-resource natural language processing (NLP).

SWETA POUDEL received her B.Tech degree in Computer Engineering from Kathmandu Engineering College, Nepal. Following the completion of her B.Tech degree, she started her professional career as a software engineer, a role in which she continues to serve to this day. She has contributed to various projects and published papers at a few conferences.

SANDESH JAIN received his MS degree in Computer Engineering from Virginia Tech and is presently engaged in pursuing a Ph.D. in the field of Computer Vision and Machine Learning at the same institution. His research focuses on multimodal computer vision, leveraging deep learning for 3D perception, and developing scalable machine learning algorithms. Currently serving as a Computer Vision Graduate Research Assistant at the Virginia Tech Transportation Institute, his work primarily revolves around 3D object detection, tracking, sensor fusion, and automated calibration for truck-based Advanced Driver Assistance Systems (ADAS). His contributions extend to projects sponsored by the Federal Motor Carrier Safety Administration (FMCSA), and he has presented research papers at conferences targeting computer vision and vehicle intelligence.

ANUM MASOOD received the B.Sc. and M.Sc. degrees in Computer Science from the COMSATS University Islamabad, Islamabad, Pakistan, and the Ph.D. degree in Computer Science and Engineering from the Shanghai Jiao Tong University, Shanghai, China. She worked as a Lecturer with the Department of Computer Science, COMSATS University Islamabad, Islamabad, Pakistan, from 2014 to 2020. She is a postdoctoral researcher with the Norwegian University of Science and Technology, Norway, and affiliated with PET Centre, St. Olav's Hospital, Trondheim, Norway. In addition, she is a Research Fellow at the Radiology Department, Harvard Medical School, Harvard University, and Boston Children's Hospital, Boston, U.S.A. She also worked as a visiting researcher with the Institute of Neuroscience and Medicine, Forschungszentrum Jülich; Institute for Cardiogenetics, University of Luebeck, Germany; and Liverpool John Moores University, Liverpool, U.K. Her research interests include medical image analysis, automated cancer detection, machine learning, and image processing.

USMAN NASEEM is a Lecturer with the School of Computing, Macquarie University, Sydney, Australia. Before pursuing his research, he worked in leading ICT companies for over ten years. His research interests include natural language processing (NLP) and computational social science, focusing on understanding human communication in social contexts and developing socially aware language technologies. He publishes and serves with the Program Committee, including the Area Chair for several top-tier venues, including ACL, EMNLP, NAACL, COLING, WSDM, Webconf, and AAAI. He also delivered several invited talks and a tutorial on recommender systems at Webconf 2023. His work in NLP has attracted attention from the World Health Organization and earned him the Nepean Blue Mountains Local Health District (NBMLHD) Board Chair's Quality Award, in 2021. He also received the prestigious IEEE Best Transactions Paper Award, in 2022.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

IOP Conference Series: Earth and Environmental Science

PAPER • OPEN ACCESS

Numerical modeling of microstructure of slag incorporating cement paste

To cite this article: Andualem Yadeta *et al* 2024 *IOP Conf. Ser.: Earth Environ. Sci.* **1326** 012071

View the article online for updates and enhancements.

You may also like

- The Present Situation and Analysis of Slag Splashing Technology of Converter in China
Tang Biao, Zhao Zhongying, Li Hongbo et al.
- Characteristics of laser-induced breakdown spectroscopy of liquid slag
Changyan DONG, , Hongxia YU et al.
- Effect of slag powder on strength development of alkali-activated recycled cement
Chenhui Zhu, Yongfan Gong and Yonghao Fang

This content was downloaded from IP address 14.139.251.98 on 01/07/2024 at 05:34

Numerical modeling of microstructure of slag incorporating cement paste

Andualem Yadeta, Pradeep Goyal, and Raju Sarkar

Department of Civil Engineering, Delhi Technological University, Delhi 110042, India

Email: andu1781@gmailcom

Abstract. This paper presents a numerical modeling approach aimed at illustrating the complicated microstructure of slag incorporating cement paste. Slag, a by-product of the steel industry, is a sustainable alternative to traditional cementitious materials, and understanding its microstructural characteristics is vital for optimizing its performance in concrete applications. In this study, an advanced computational technique named the CEMHYD3D model was employed to simulate the microstructure of slag incorporating cement paste mixtures. The modeling method involves the discretization of the microstructure and the incorporation of complex hydration reactions. Through this approach, the evolution and microstructural development within the slag-cement paste system over time has been predicted. The microstructural properties of the blend were simulated based on various slag replacement levels and the result agreed with the experimental results. The simulation provides understandings into the interaction between slag particles and the surrounding cementitious matrix, shedding light on the development of the microstructural properties of the composite material. In summary, this paper advances the understanding of the microstructural development of slag-cement paste system through rigorous numerical modeling. The insights gained from these simulations have significant implications for optimizing the use of slag as a sustainable and environmentally friendly component in concrete production.

1. Introduction

Supplementary cementitious materials (SCMs) such as fly ash (FA), silica fume (SF), and ground granulated blast furnace slag (GGBFS), which is also known as slag, are commonly used in concrete mixes due to their environmental and economic benefits. When added to Portland cement, SCMs change the characteristics of a cementitious material through either their pozzolanic activity and/or hydraulic activity. Slag is among the most commonly used SCMs and is produced from iron blast-furnace slag during the reduction of iron ore at a temperature of about 1500ºC.

Slag has substantial activity during the hydration and microstructural development of cementitious materials [1], [2]. It is further described that the substitution of slag into Portland cement improves the workability and hydration of cement [1], [3], [4]. Cement particles are composed of tri-calcium silicate (C_3S), di-calcium silicate (C_2S), tri-calcium aluminate (C_3A), and tetra-calcium aluminate ferrite (C_4AF) mineral phases. The effect of slag on the microstructural properties of cement has been experimentally explored in previous studies [4]–[7], and it has been explained that it changes the pore structure of cementitious materials.

Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

Materials characterization techniques like scanning electron microscopy (SEM) [8] and X-ray diffraction (XRD) [9] are usually used to experimentally determine the microstructural properties of cement-based materials. However, these techniques are time-consuming and labor-intensive. Due to the current advancement in computational science, various computer models were created to predict the various microstructural properties of cementitious materials. Based on the hydration process for cementitious systems that use Portland cement, a number of models, such as CEMHYD3D [10], HYMOSTRUC3D [11], [12], μic [13] and DuCOM [14] were developed to simulate the microstructure of cement-based materials.

The microstructure of SCMs incorporating cement-based materials is more complex than that of cementitious systems involving only Portland cement, owing to the mineral reactions involved in each cement. The mineral reactions during the hydration of cements involving SCMs such as slag, FA, and limestone are considered distinct from those involving regular Portland cement, and some reactions are considered by looking at the concentration of calcium hydroxide and free water. The hydration of the minerals contained in each individual cement in cement paste affects the microstructure of cement paste. Even though numerous models were developed in previous studies, they predict the hydration and microstructural characteristics of the Portland cementitious system. A model to predict the microstructural development of slag incorporating cement has not yet been developed. Therefore, there is a need to develop a prediction model to simulate microstructural development in slag incorporating cement paste.

2. Modeling Approach

The National Institute of Standards and Technology originally developed a modeling software named CEMHYD3D to simulate the microstructure of cement [15]. The microstructure is represented by a grid of volume pixels, each of which represents a certain hydration phase in this model. To accurately represent the distributions of phases in the cementitious system, the model simulates the cement particles based on real cement images and particle size distributions. In order to simulate the microstructure of slag-incorporating cement paste, the extended version of the CEMHYD3D model has been employed in this study.

The chemical effects of the minerals in cement on hydration reactions and microstructure development need to be included in modeling. Calcium hydroxide (CH) and calcium-silicate-hydrate (C-S-H) are formed when C_3S and C_2S are hydrated with water. As long as C_3A or C_4AF aluminate clinker phases are present, aluminate is absorbed by C-S-H which produces a C-A-S-H product. If no remaining aluminate reactants are left, C_3S and C_2S reactions further yield CSH without aluminum substitution. The processes that the minerals in Portland cement go through during the hydration stage can be explained with equations (1)-(4), as described in [16]–[19].

$$2C_3S + 6H \rightarrow C_3S_2H_3 + 3CH \quad (1)$$

$$2C_2S + 4H \rightarrow C_3S_2H_3 + CH \quad (2)$$

$$C_3A + C\overline{S}H_2 \rightarrow 10H + C_4A\overline{S}H_{12} \quad (3)$$

$$C_4AF + 2CH \rightarrow 10H + C_6AFH_{12} \quad (4)$$

For a slag incorporating cement, there are five major hydration products that form when slag is added to cement: C-S-H, C-S-A-H, hydrotalcite (M_5AH_{13}), ettringite ($C_6A\overline{S}_3H_{32}$), and AFm (C_4AH_{13}) phases are [2]. While both slag and Portland cement produce similar hydration products, part of the portlandite produced when cement clinker hydrates is actually consumed when slag is hydrated [20]–[22]. Nevertheless, a high rate of slag reaction requires a lower w/c ratio [16]. The reaction of slag during hydration is appropriately summarized by the approximate critical values shown in table 1, as explained in [23], based on a study conducted on blended cements.

Table 1. Reaction of slag [23]

Quantity measured	Values (g/g of slag)
Calcium hydroxide	0.22
Chemically-bound water	0.30
Gel water	0.15

In accordance with the stoichiometry and hydration model of slag reaction [23], the mass of calcium hydroxide, the chemically-bound water, and the gel water may all be computed with equations (5)-(7) in which the hydration of slag-incorporating cements can be described.

$$CH = RCH_{ce}\, x\, C_0\, x\, \alpha - RCH_{sg}\, x\, \alpha_{SG}\, x\, P \tag{5}$$

$$W_{cap} = W_0 - 0.4\, x\, C_0\, x\, \alpha - RCW_{sg}\, x\, \alpha_{sg} xP - RPW_{sg} x\, \alpha_{sg}\, x\, P \tag{6}$$

$$W_{cbw} = v\, x\, C_0\, x\, \alpha + RCW_{sg}\, x\, \alpha_{sg}\, x\, P \tag{7}$$

where CH, W_{cap}, and W_{cbw} are the mass of calcium hydroxide, capillary water, and chemically-bound water, respectively. *sg* is the rate of the slag reaction, C_0 and W_0 are the initial percentages by mass of cement and water, respectively, in the mixture; P is the mass of the slag in the mix; RCH_{ce} and RCH_{sg} are the mass of calcium hydroxide formed in the hydration of cement and the slag reaction in that order; where as RCW_{sg} and RPW_{sg} are the mass of chemically-bound water and the gel water in the slag reaction, respectively.

3. Model validation

A Type I Cement, or an ordinary Portland cement (OPC), as per the standard stipulated in ASTM C 150, has been used as a reference cement in this study. Additionally, a slag confirmed to meet ASTM C 989 standard requirements is used both in the experiment and the model. The particle size distribution (PSD) of each cement plays a crucial role in the performance of cement-based materials. Both slag and OPC consist of a range of particle sizes, and understanding their distribution is essential for predicting the microstructural properties of cementitious composites. The PSD of each cement slag was obtained using a laser diffraction (LD) technique as presented in figure 1.

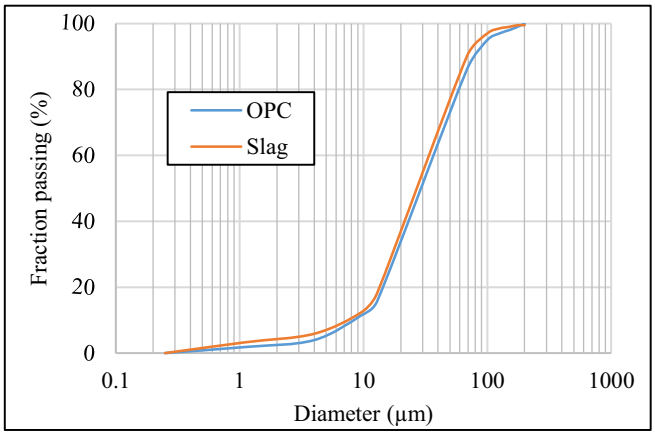


Figure 1. Particle size distribution of OPC and slag used in the model

The oxide composition of each cement is another fundamental aspect when investigating the microstructural properties of cementitious materials. These oxides significantly influence the chemical reactions, hydration kinetics, and ultimately the resulting microstructure of the paste. The oxides are represented with standard abbreviations in the reactions: C stands for CaO, S for SiO_2, A for Al_2O_3, F for Fe_2O_3, M for MgO, and H for H_2O. Slag is known for its high silica (SiO_2) content, which imparts pozzolanic properties to the material. Silica-rich slag can react with calcium hydroxide (CH) produced during cement hydration to form calcium silicate hydrate (C-S-H) gel. This C-S-H gel is the primary binding phase in cement paste and contributes to its strength and durability. The oxide composition of slag determines its reactivity, with higher a silica content generally leading to enhanced pozzolanic activity and better microstructural development in the paste. The oxide compositions presented in table 2 were used in the model to simulate the microstructural development of slag incorporating cementitious system. The simulation was done by considering various replacement levels of slag with cement.

Table 2. Oxide composition of OPC and slag used in the model (mass%)

Material	SiO_2 %	CaO %	Al_2O_3 %	SO_3 %	MgO %	Fe_2O_3 %	K_2O %	Na_2O %
OPC	19.10	62.38	5.35	3.34	2.37	2.95	1.05	0.05
Slag	40.23	35.04	11.18	6.80	2.32	2.05	0.43	0.29

The porosity of the paste was measured using a mercury intrusion porosimetry (MIP) technique based on the idea that mercury intrusion volume into a porous medium is dependent on applied pressure.
If the pore shape is assumed to be cylindrical, the Washburn equation [24] can be used to relate the pore size or radius r (μm) to the applied pressure P (MPa) as expressed in equation (8):

$$r = -\frac{4\gamma cos\theta}{P} \tag{8}$$

where γ is the surface tension and θ is the contact angle between the liquid and the solid.

For porosity analysis, cement paste samples were prepared from OPC with slag substitution levels of 10%, 20%, 30%, and 40% with a w/c ratio of 0.45. The samples were formed in small tube molds. The samples were demolded after 24 hours and cured for 180 days at a standard temperature of 20°C under a saturated condition. The porosity of the blended paste samples was determined using the MIP method. Then, using the same data used in experiments, virtual cement paste samples were created to simulate the pore structures with computer models.

A modeling software called CEMHYD3D, which is available in the virtual cement and concrete testing laboratory (VCCTL) from the open source, was employed to simulate the microstructural development of slag-blended cement. As the code in the CEMHYD3D model allows us to adjust the model, the reaction of slag has been incorporated into the program. Virtual samples of cement pastes were created with a constant w/c of 0.45 at a slag replacement level of 10%, 20%, 30%, and 40% by mass of cement. Then the pore structure was simulated using the extended version of the CEMHYD3D model. The model prediction has been validated with measured values from experiments based on the standard test method of ASTM C186. Previously, experimental studies were conducted by [5], [25], [26]. In slag-blended cement, the amount of CH and the rate of reaction of mineral components determine the rate of hydration of the cement [1], [5]. This proves that the hydration product of slag usually develops in close proximity to the slag particles [27]–[29].

4. Simulation of pore structure of slag incorporating cement paste

Simulating the pore structure of slag incorporating cement paste with various slag replacement levels is crucial for understanding how slag affects the microstructure and, consequently, the performance of the composite material. The pore structure of this cement paste was simulated using slag replacement levels of 0%, 10%, 20%, 30%, and 40% while maintaining a constant w/c of 0.45, with a focus on the predicted pore

characteristics at a curing age of 180 days as separately presented in a, b, c, and d in figure 2. The predicted porosity agreed with the measured values. The prediction from the model shows that the porosity of the paste increases as the percentage of slag substitution increases during the particular curing age of 180 days in this study. This potentially increases the coefficient of transport of ions or moisture in the composite. Microstructures from a scale of nanometers to micrometers govern the permeability of cement-based materials [30], [31]. That is why microstructural properties are important to exhibit excellent durability properties in cementitious materials.

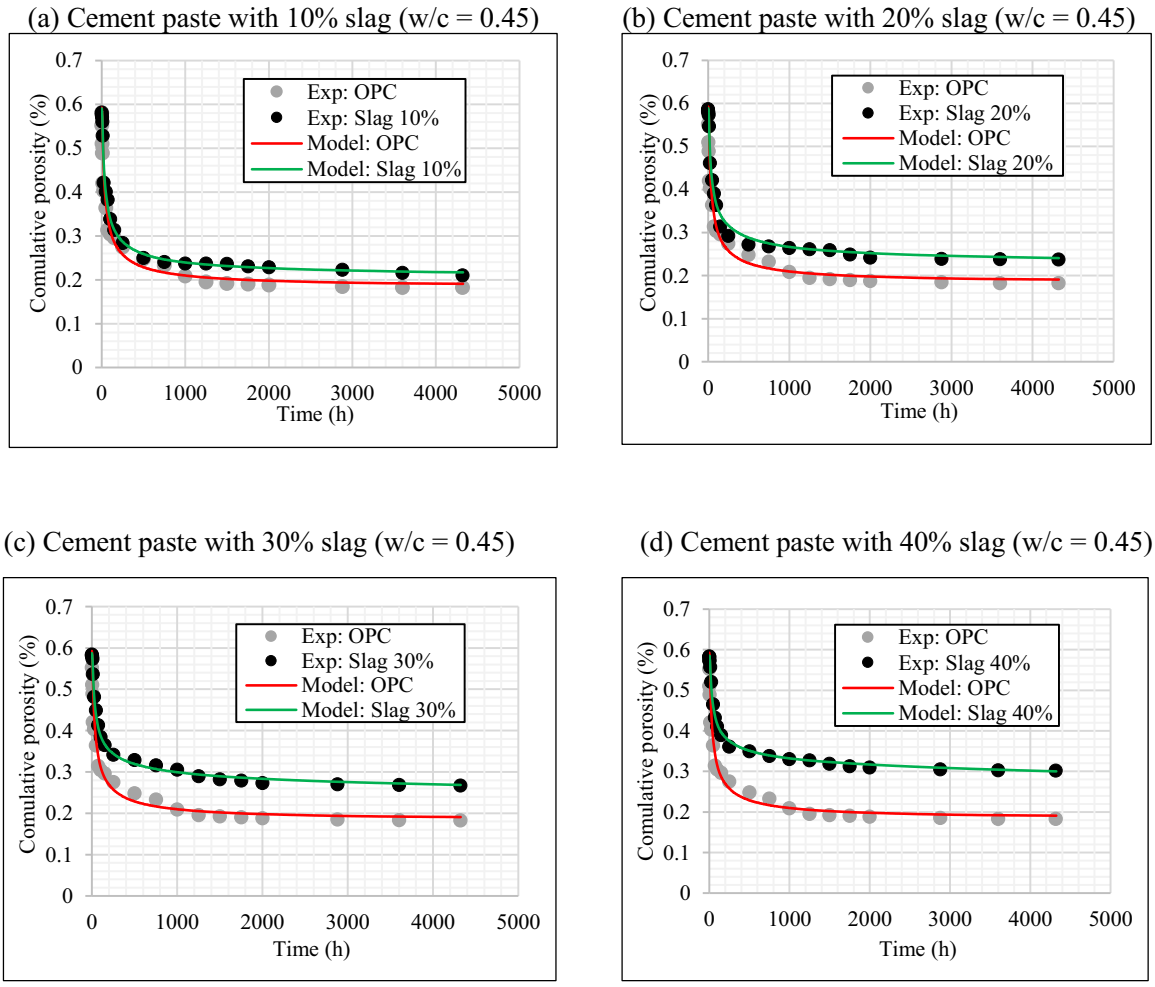
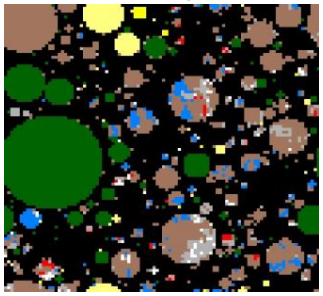


Figure 2. Simulated pore structure of slag incorporating cement pastes with various replacement levels for the curing age of 180 days

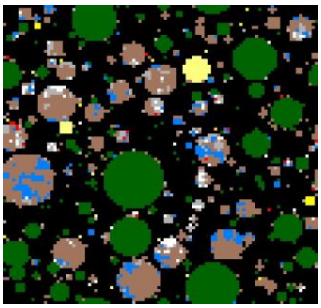
Based on the volume-pixel modeling approach, detailed microstructural properties can be obtained with simulation [32]. The simulation from the model also indicates that the outer hydrate layer expands while the volume of small capillary pores in the system decreases as the rate of hydration increases. Pore refinement is usually studied when SCMs are substituted in cement and is further used to explain the reduction of permeability and diffusivity of cementitious composites [33]. The distribution mode for the size of small capillary pores is similar across all samples of slag incorporating cement pastes, as shown in figure 3 below.

(a) Microstructure of blended cement paste with 10% slag (w/c = 0.45)

(b) Microstructure of blended cement paste with 20% slag (w/c = 0.45)



(c) Microstructure of blended cement paste with 30% slag (w/c = 0.45)



(d) Microstructure of blended cement paste with 40% slag (w/c = 0.45)

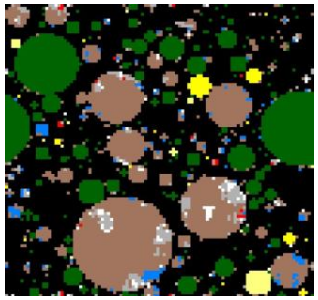


Figure 3. Simulated microstructure of slag incorporating cement pastes from the CEMHYD3D model with various slag substitution levels. The colors in the images indicate different phases. Brown is C_3S, blue is C_2S, grey is C_3A, white is C_4AF, yellow is gypsum, red is K_2SO_4, green is free lime, aqua is free silica, pink is periclase, and black is pore structure.

5. Conclusion

This study introduced the simulation of the microstructural development of slag incorporating cement paste using computer models. The simulation was done to investigate the pore structure of the blended cement system with slag replacement levels of 10%, 20%, 30%, and 40% made with a w/c ratio of 0.45 cured in a saturated condition at a standard temperature of 20°C. The prediction values agreed with the measured values, and it was successful in predicting the microstructural properties of slag incorporating cement paste. The prediction was done only to investigate the pore structure of slag incorporating cement pastes. Further studies are recommended to simulate and validate the model for other SCMs incorporating cement pastes at various w/c, varying curing temperatures, and saturation conditions to determine the various microstructural properties.

6. References

[1] B. Kolani, L. Buffo-Lacarrière, A. Sellier, G. Escadeillas, L. Boutillon, und L. Linger, „Hydration of slag-blended cements“, *Cem Concr Compos*, Bd. 34, Nr. 9, S. 1009–1018, Okt. 2012, doi: 10.1016/j.cemconcomp.2012.05.007.

[2] W. Chen und H. J. H. Brouwers, „The hydration of slag, part 2: Reaction models for blended cement“, *J Mater Sci*, Bd. 42, Nr. 2, S. 444–464, Jan. 2007, doi: 10.1007/s10853-006-0874-1.

[3] S. W. Yoo und S. J. Kwon, „Effects of cold joint and loading conditions on chloride diffusion in concrete containing GGBFS“, *Constr Build Mater*, Bd. 115, S. 247–255, Juli 2016, doi: 10.1016/j.conbuildmat.2016.04.010.

[4] Y. C. Choi, J. Kim, und S. Choi, „Mercury intrusion porosimetry characterization of micropore structures of high-strength cement pastes incorporating high volume ground granulated blast-furnace slag“, *Constr Build Mater*, Bd. 137, S. 96–103, Apr. 2017, doi: 10.1016/j.conbuildmat.2017.01.076.
[5] Y. Zhao, J. Gao, Z. Xu, S. Li, X. Luo, und G. Chen, „Long-term hydration and microstructure evolution of blended cement containing ground granulated blast furnace slag and waste clay brick“, *Cem Concr Compos*, Bd. 118, Apr. 2021, doi: 10.1016/j.cemconcomp.2021.103982.
[6] J. Xing, Y. Zhao, J. Qiu, und X. Sun, „Microstructural and mechanical properties of alkali activated materials from two types of blast furnace slags“, *Materials*, Bd. 12, Nr. 13, Juli 2019, doi: 10.3390/ma12132089.
[7] J. Y. Lee, J. S. Choi, T. F. Yuan, Y. S. Yoon, und D. Mitchell, „Comparing properties of concrete containing electric arc furnace slag and granulated blast furnace slag“, *Materials*, Bd. 12, Nr. 9, 2019, doi: 10.3390/ma12091371.
[8] H. Binici, H. Temiz, und M. M. Köse, „The effect of fineness on the properties of the blended cements incorporating ground granulated blast furnace slag and ground basaltic pumice“, *Constr Build Mater*, Bd. 21, Nr. 5, S. 1122–1128, Mai 2007, doi: 10.1016/j.conbuildmat.2005.11.005.
[9] M. Zhang, „Pore-scale modelling of relative permeability of cementitious materials using X-ray computed microtomography images“, *Cem Concr Res*, Bd. 95, S. 18–29, Mai 2017, doi: 10.1016/j.cemconres.2017.02.005.
[10] D. P. Bentz, „CEMHYD3D“:, Gaithersburg, MD, 2005. doi: 10.6028/NIST.IR.7232.
[11] K. Van Breugel, „5) 0 0 0 1 7-8 NUMERICAL SIMULATION OF HYDRATION AND MICROSTRUCTURAL DEVELOPMENT IN HARDENING CEMENT-BASED MATERIALS (I) Theory“, 1995.
[12] P. Apa) Gao, G. ; Ye, J. ; Wei, und Q. Yu, „Extension of the Hymostruc3D model for simulation of hydration and microstructure development of blended cements“, 2019.
[13] S. Bishnoi und K. L. Scrivener, „μic: A new platform for modelling the hydration of cements“, *Cem Concr Res*, Bd. 39, Nr. 4, S. 266–274, Apr. 2009, doi: 10.1016/j.cemconres.2008.12.002.
[14] K. Maekawa, T. Ishida, und T. Kishi, „Multi-scale Modeling of Concrete Performance Integrated Material and Structural Mechanics“, 2003.
[15] D. P. Bentz, „CEMHYD3D“:, Gaithersburg, MD, 2005. doi: 10.6028/NIST.IR.7232.
[16] V. G. Papadakis, „Effect of supplementary cementing materials on concrete resistance against carbonation and chloride ingress“, 2000.
[17] X. Lu, Z. Ye, S. Wang, P. Du, C. Li, und X. Cheng, „Study on the preparation and properties of belite-ye'elimite-alite cement“, *Constr Build Mater*, Bd. 182, S. 399–405, Sep. 2018, doi: 10.1016/j.conbuildmat.2018.06.143.
[18] M. H. Maciel, G. S. Soares, R. C. de O. Romano, und M. A. Cincotto, „Monitoring of Portland cement chemical reaction and quantification of the hydrated products by XRD and TG in function of the stoppage hydration technique“, *J Therm Anal Calorim*, Bd. 136, Nr. 3, S. 1269–1284, Mai 2019, doi: 10.1007/s10973-018-7734-5.
[19] C. Pignat, P. Navi, und K. Scrivener, „Simulation of cement paste microstructure hydration, pore space characterization and permeability determination“, *Materials and Structures/Materiaux et Constructions*, Bd. 38, Nr. 278, S. 459–466, Mai 2005, doi: 10.1617/14355.
[20] K. Luke und E. Lachowski, „Internal composition of 20-year-old fly ash and slag-blended ordinary portland cement pastes“, *Journal of the American Ceramic Society*, Bd. 91, Nr. 12, S. 4084–4092, Dez. 2008, doi: 10.1111/j.1551-2916.2008.02783.x.
[21] R. Taylor, I. G. Richardson, und R. M. D. Brydson, „Composition and microstructure of 20-year-old ordinary Portland cement-ground granulated blast-furnace slag blends containing 0 to 100% slag“, *Cem Concr Res*, Bd. 40, Nr. 7, S. 971–983, Juli 2010, doi: 10.1016/j.cemconres.2010.02.012.
[22] J. J. Thomas, A. J. Allen, und H. M. Jennings, „Density and water content of nanoscale solid C-S-H formed in alkali-activated slag (AAS) paste and implications for chemical shrinkage“, *Cem Concr Res*, Bd. 42, Nr. 2, S. 377–383, Feb. 2012, doi: 10.1016/j.cemconres.2011.11.003.

[23] K. Maekawa, T. Ishida, und T. Kishi, *Multi-scale Modeling of Structural Concrete*, 1st Edition. New York: Taylor & Francis, 2009.
[24] R. Kumar und B. Bhattacharjee, „Porosity, pore size distribution and in situ strength of concrete“, *Cem Concr Res*, Bd. 33, Nr. 1, S. 155–164, Jan. 2003, doi: 10.1016/S0008-8846(02)00942-0.
[25] P. T. Durdziński, M. Ben Haha, M. Zajac, und K. L. Scrivener, „Phase assemblage of composite cements“, *Cem Concr Res*, Bd. 99, S. 172–182, Sep. 2017, doi: 10.1016/j.cemconres.2017.05.009.
[26] M. Hlobil, V. Šmilauer, und G. Chanvillard, „Micromechanical multiscale fracture model for compressive strength of blended cement pastes“, *Cem Concr Res*, Bd. 83, S. 188–202, Mai 2016, doi: 10.1016/j.cemconres.2015.12.003.
[27] C. Çetin, S. T. Erdołan, und M. Tokyay, „Effect of particle size and slag content on the early hydration of interground blended cements“, *Cem Concr Compos*, Bd. 67, S. 39–49, März 2016, doi: 10.1016/J.CEMCONCOMP.2015.12.001.
[28] X. Y. Wang, „Analysis of hydration kinetics and strength progress in cement–slag binary composites“, *Journal of Building Engineering*, Bd. 35, S. 101810, März 2021, doi: 10.1016/J.JOBE.2020.101810.
[29] X. Feng, E. J. Garboczi, D. P. Bentz, P. E. Stutzman, und T. O. Mason, „Estimation of the degree of hydration of blended cement pastes by a scanning electron microscope point-counting procedure“, *Cem Concr Res*, Bd. 34, Nr. 10, S. 1787–1793, Okt. 2004, doi: 10.1016/J.CEMCONRES.2004.01.014.
[30] S. Xu, G. C. Simmons, T. S. Mahadevan, G. W. Scherer, S. H. Garofalini, und C. Pacheco, „Transport of water in small pores“, *Langmuir*, Bd. 25, Nr. 9, S. 5084–5090, Mai 2009, doi: 10.1021/la804062e.
[31] S. Bishnoi und J. W. Bullard, „Microstructure models of cement: their importance, utility, and current limitations“, *RILEM Technical Letters*, Bd. 6, S. 188–195, März 2021, doi: 10.21809/rilemtechlett.2021.135.
[32] Q. H. Do, S. Bishnoi, und K. L. Scrivener, „Numerical Simulation of Porosity in Cements“, *Transp Porous Media*, Bd. 99, Nr. 1, S. 101–117, Aug. 2013, doi: 10.1007/s11242-013-0176-4.
[33] E. Berodier und K. Scrivener, „Evolution of pore structure in blended systems“, *Cem Concr Res*, Bd. 73, S. 25–35, Juli 2015, doi: 10.1016/J.CEMCONRES.2015.02.025.

Microsystem Technologies
https://doi.org/10.1007/s00542-024-05705-z

TECHNICAL PAPER

Numerical modelling for triple hybrid gate optimization dielectric modulated junctionless gate all around SiNWFET based uricase and ChO_X biosensor

Rishu Chaujar[1] · Mekonnen Getnet Yirak[1,2]

Received: 6 October 2023 / Accepted: 4 June 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract
In this manuscript, a numerical model based on the electric field, threshold voltage, sub-threshold current, and electrostatic potential in cylindrical coordinates using Poisson's equation for triple hybrid metal (THM) gate dielectric modulated junctionless silicon-nanowire gate all around FET based uricase and ChO_X biosensor was developed at 40 nm technology (20 nm gate length) to study different gate engineering optimization effects on the performance of the proposed device. The results of the ATLAS-3D TCAD" device simulator agreed with a derived analytical model. Three types of gate optimization (gate engineering) are denoted by M_ϕ (4.86, 4.96 and 4.50 eV), O_ϕ (4.96, 4.86 and 4.50 eV), and Q_ϕ (4.86, 4.50 and 4.96 eV) each have three different metal work-function, including uricase and cholesterol oxidase (ChO_X) biomolecules have been coated in the nanocavity to determine their impact on the device performance and also, the effect of nanogap cavity length on the proposed device was examined taking numerous simulations. Our findings conclude that nanocavity coated with ChO_X dielectric and having tunable work-function optimized at "O" signifies better output results in the device sensitivity, shifting threshold voltage, switching ratio, transconductance, intrinsic voltage gain, and device efficiency. For instance, the switching ratio in the case of ChO_X biomolecule for M, O, and Q gate optimizations are 5.22×10^5, 1.36×10^6, and 2.18×10^4, respectively. We conclude that the proposed devices with optimizing gate work function at "O" suggest new opportunities for future ultra-large-scale integration (ULSI) development to achieve highly efficient device performance.

1 Introduction

Polysilicon with a double-type doping profile (P or N) has conventionally been used as conductive gate materials for bulk MOSFETs (Sachdeva et al. 2018; Avila et al. 2016; Dutta and Kundu 2012). As CMOS technology continuously scales down to ultra-small device size (Chaujar et al. 2008; Madan and Rishu 2016; Hossain et al. 2017), metal gate electrodes need to be presented to alleviate harmful effects of the doped polysilicon gate, like high parasitic resistance, gate electrode depletion, and power consumption (Dutta and Kundu 2012; Sze 2014). A depletion layer near the polysilicon/oxide interface is developed due to reduced active dopant levels in the polysilicon gate when the device is biased in strong inversion, which in turn results in dishonored device characteristics (Wang et al. 2008), the so-called polysilicon depletion or poly-depletion effect (Wenga et al. 2013) and leads to capacitance and drain current value reduction in addition to the undesirable variation of the threshold voltage due to unexpected electronic circuit output results and polysilicon depletion effect (Kale et al. 2018; Peesa and Panda 2021). To reduce those polysilicon depletion effects like threshold voltage roll-off, parasitic gate resistance, and short channel effects (SCEs) (Kumar et al. 2018), optimizing gate electrodes with tunable (high metal) work functions instead of the polysilicon gate is crucial (Chaujar et al. 2009; Houl et al. 2004; Ana and Din 2011) under the high-k dielectric gate oxide (Chau et al. 2004; Peesa and Panda 2022). Also, different researchers inspect various approaches like surrounded

✉ Rishu Chaujar
chaujar.rishu@dtu.ac.in

Mekonnen Getnet Yirak
mekonnengetnet01@gmail.com

[1] Applied Physics Department, Delhi Technological University, Delhi, India

[2] Physics Department, Debre Tabor University, Debre Tabor, Ethiopia

Published online: 18 June 2024

gate metals (Siddik et al. 2017), fully silicide gates, multi-gate device structures (Avila et al. 2016; Goel et al. 2018; Reddy and Panda 2022), and composite metal gates with tunable work function (Reddy and Panda 2022; Maszara 2005) to curtail polysilicon depletion effects. Different researchers' evidence approved that metal gate electrodes with high work function used as gate electrodes (interface) to alleviate problems caused by a polysilicon gate in the enhancement of ultra-small scall devices (Peesa and Panda 2022; Reddy and Panda 2022; Liu et al. 2005). Since "gate resistance–capacitance delay becomes a critical concern in designing electronic circuits with very short and thin gate oxides" (Kumar et al. 2018; Sjöblom 2006) under small-scale dimensions. Unlike polysilicon electrodes, which have two work functions only (N or P, type), electrodes(metal) can set almost any work function for CMOS devices (Reddy and Panda 2022). Currently, different researchers have proved that the tuned work function lies between 4.50 eV and 5.00 eV and effectively operates for CMOS transistors (Chau et al. 2004; Biswal et al. 2019; Rechem et al. 2009) under high-k dielectric gate oxide in ultra-small scall device development since tunable electrode work-function is crucial in determining the appropriate threshold voltage values and suppresses short channel effects (Biswal et al. 2019; Rony et al. 2016; Kumar et al. 2020). In ultra-small scale device architecture, not only the polysilicon gate but also the fabrication process for NWFET is difficult due to the severe doping necessities at (drain, channel, and source) in the area of nano dimension gaps (Rana et al. 2020; Parihar and Kranti 2015), and the gate has less control over the drain current. These challenges in the fabricating process of NWFETs are addressed by developing junctionless (JL) FETs since there are no doping concentration gradients between the source and drain(in the device) (Rana et al. 2020; Georgiev et al. 2017; Colinge et al. 2011; Shreya et al. 2020) even if the device is heavily doped. The doping profile should be uniform in the source/drain and substrate regions to eliminate the junction from the transistor and make it junctionless. Because basic "physical properties such as drain-induced barrier lowering (DIBL) sub-threshold swing, parasitic resistance, transconductance, Ion/Ioff ratio, and other physical properties have been improved in the junctionless phenomenon" (Kumar et al. 2020; Pratikhya Raut and Nanda 2023).

Furthermore, the gate-all-around (GAA) architecture allows comprehensive drain current gate control (Siddik et al. 2017; Pradhan et al. 2015). The flow of electrons through the wire is governed by the voltage supplied across the gate terminal. The GAA JL-NWFET utilizes NW of silicon surrounded by an electrically isolated metal electrode that serves as a gate.

Advanced device structures like multi-metal gate and junctionless gate-all-around (GAA) nanowire FETs have received crucial attention due to better scalability and amended capacitive coupling in gate-to-channel regions (Reddy and Panda 2022; Nam et al. 2016). However, optimization of gate electrode work function on the multi-gate uricase and ChO_X biosensor devices has not yet been well investigated in the ultra-small scale device dimension (Matsukawa et al. 2008). Different works explore that SiNWFET-based uricase and ChO_X biosensors have low cost, low power consumption, and easy-to-use on-site monitoring, even by non-medical staff, to achieve effective surveillance of blood metabolites at home (Zhang and Liu 2015; Li et al. 2013; Reddy and Panda 2022). The prime origins of mortality and disability linked to abnormal glucose concentration problems are caused by an imbalance of blood metabolites and diabetes (Xuan et al. 2018; Dey and Raj 2013). So, controlling and managing the amounts of uric acid, cholesterol, and glucose concentration in human fluids is essential for living a healthy life (Dey and Raj 2013). Uricase and ChO_X biosensors are especially important in clinical diagnostics, mainly in monitoring gout and hyperuricemia (Pratap et al. 2018). Gout is a kind of arthritis that occurs when uric acid crystals accumulate in joints, producing inflammation and intense discomfort. Hyperuricemia, or high amounts of uric acid in the blood, can cause gout and other health problems such as kidney stones and cardiovascular disease (Piermarini et al. 2013; Pasalic et al. 2012).

As a result, SiNWFET-based uricase and ChO_X biosensors have been actively investigated to detect cholesterol (Piermarini et al. 2013; Soylemez et al. 2015), uric acid (Piermarini et al. 2013; Pasalic et al. 2012), and glucose concentration due to their high sensitivity, simplicity, and low cost (Pasalic et al. 2012; Getnet and Chaujar 2022; Rahman et al. 2017), and their higher specific surface area for enzyme immobilization than planar electrodes, which improves sensitivity (Ambhorkar et al. 2018; Gao et al. 2019) when uricase and ChO_X biomolecules are immobilized on the nanogap cavity in the form of an aqueous solution. Since nanometer-sized molecules can be detected using NWFET-based sensors (Dey and Raj 2013; Pratap et al. 2018; Piermarini et al. 2013). They provide improved sensitivity, electrical detection without labels, and mass production (Yoon et al. 2020; Huang et al. 2018; Lu et al. 2014). These biosensors have several advantages over traditional approaches, including portability, real-time monitoring, and lower sample volume requirements (Getnet Yirak and Chaujar 2020). Understanding the importance of uricase and ChO_X biosensors is critical for developing healthcare and improving patient outcomes (Gao et al. 2019; Lee et al. 2019). It is possible to synthesize/characterize

biomolecules like uricase and ChOx biomolecules in the form of nanometer size using transmission electron microscopy (TEM) and X-ray photoelectron spectroscopy (XPS) instruments (Kumar et al. 2020; Yoon et al. 2020; Yan et al. 2020; Elaty and Fakhry 2019) to develop device sensitivity and selectivity. Therefore, the sensing device should have a smaller feature size to offer appropriate sensing capabilities for such small biological or chemical species. The size of a variety of biological species is depicted in Fig. 1.

In this paper, we have proposed to study the optimization of different gate electrode work-function on the electrical output properties, such as drain current (I_{DS}), subthreshold swing (S), transconductance (g_m), threshold voltage(V_{th}), surface potential, device efficiency (g_m/I_{DS}), output conductance (g_d), switching ratio (I_{ON}/I_{OFF}), intrinsic voltage gain (g_m/g_d) by modeling triple hybrid metal gate dielectric modulated junctionless gate all around SiNWFET based uricase and ChO_X biosensor.

1.1 Device fabrication process

In this study, the Triple JL-GAA-SiNW FET device architecture is developed using Silvaco's ATLAS-3D TCAD tool. Its proficient modeling capabilities, support for multi-physics simulation, and user-friendly interface make it ideal for correctly anticipating biosensor performance and optimizing design parameters (Software 2023; Sanjay and Vohra 2020).

In the fabrication process, silicon nanowires can be fabricated utilizing the top-down method, using an e-beam lithography-specified design pattern (Pott et al. 2008; Marchi et al. 2014). Different researchers choose the top-down approach for fabricating SiNWFET-based biosensors. It enables large-scale manufacturing of nanowires without the need for complex pregrown nanowire transfer procedures on the substrate (Reddy and Panda 2022; Pott et al. 2008). Further, a fabrication flowchart of the proposed device is also placed for reference (Fig. 2).

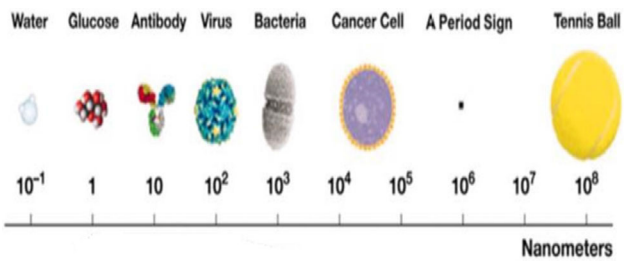


Fig. 1 Illustrates the size of biological species for characterization and electrical detection with NWFET-based biosensors (Elaty and Fakhry 2019)

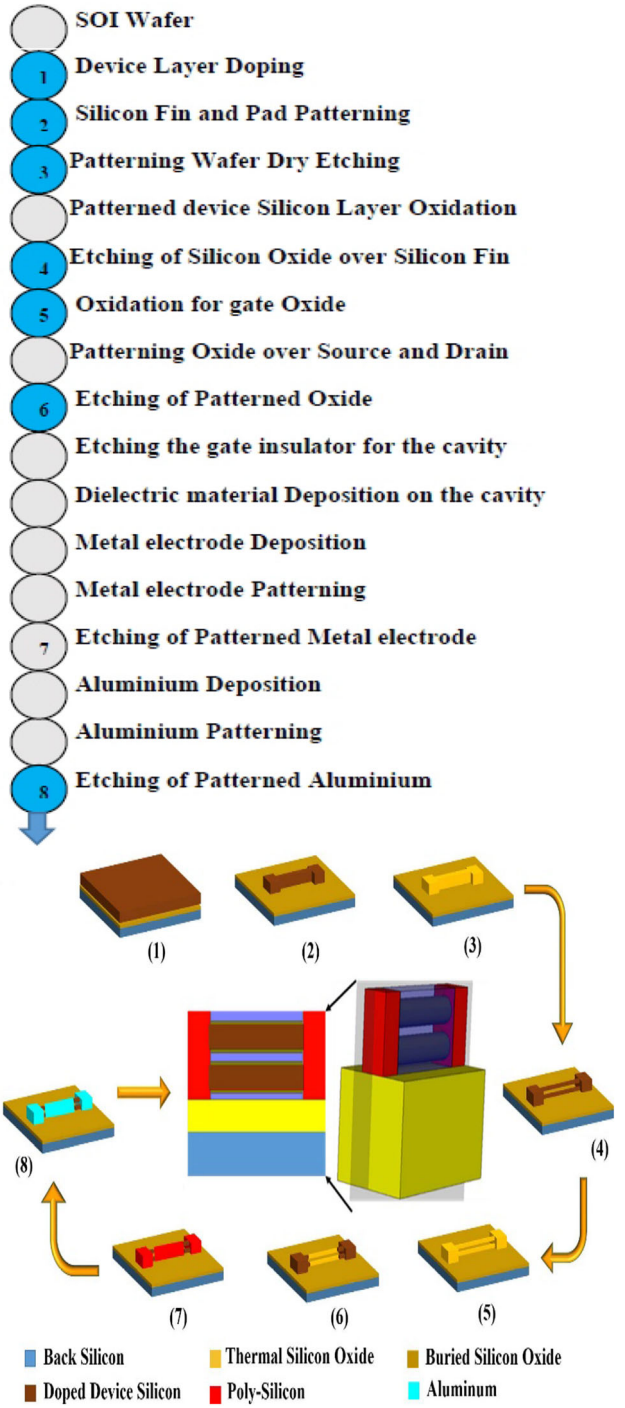


Fig. 2 Describes the fabrication process flowchart and schematic view for the GAA JL-NWFET fabrication process adapted (Biswal et al. 2019)

All these parameters and approaches will be applied to the work of Rana et al. (2020) for the possible fabrication process of the proposed device.

1.2 Device structure and simulation tools

The designed structure for n-type triple hybrid metal gate dielectric modulated junctionless gate all around NWFET-based uricase and ChO_X biosensor has been proposed, as shown in Fig. 3. Here L_1(6nm), L_2(8 nm) L_3(6 nm) are the lengths of the gate $\mathbf{G_1(\Phi_1)}$, $\mathbf{G_2(\Phi_2)}$ and $\mathbf{G_3(\Phi_3)}$ respectively and L_4 (10 nm) are the length of the nanogap cavity and silicon dioxide (SiO_2) which are near to drain and source end, respectively, where $(L_{ch} = L_1 + L_2 + L_3 = 20nm)$ is the length of hafnium oxide dielectric (HfO_2) and an interface layer oxide (SiO_2). T_{high-K}, T_{bio} and T_{OX} are the thickness of hafnium oxide ($\in = 25$), biomolecule cavity, and interface ($SiO_2, \in = 3.9$) gate oxides, respectively, and 2R is the channel's diameter (silicon film thickness). The target biomolecules (Uricase $\in = 1.54$ and $ChO_X, \in = 3.30$) are detected in the cavity region and assumed to be immobilized and interact with the cavity uniformly (Reddy and Panda 2022; Ajay et al. 2015). Doping concentration $N_D = 1x10^{19}cm^{-3}$ is applied uniformly for TM-GAA-JL-NWFET from source to drain through the channel regions to study the impacts of gate work-function variation. The proposed device calibration output result has been authenticated with the experimental results of papers (Im et al. 2007).

Several simulations were carried out using the ATLAS-3D device simulator tool to analyze the electrical properties of dielectric modulated triple hybrid metal GAA JL-NWFETs with different gate engineering, including different biomolecules, like uricase, cholesterol oxidase (ChO_X), and other biomolecules along the materials specified in Table 1. We incorporate carrier recombination and generation models for carrier mobility in the simulation. Drift–diffusion or ballistic transport models are used to simulate the movement of charge carriers (electrons or holes) inside the SiNWFETs (Sanjay and Vohra 2020). Concentration-dependent mobility (CONMOB) (Software 2023) and bandgap narrowing (BGN) models are applied to verify doping versus mobility and high channel doping profile (for doping profile is more significant than $10^{18}cm^{-3}$). Ignoring impact ionization effect, Boltz–man transport and Shockley–Read–Hall (SRH) models account for the recombination of minority carriers within the semiconductor (Software 2023). The Lombardi CVT mobility model accounted for field-dependent mobility in perpendicular/parallel (Software 2023) at room temperature (300 K). Furthermore, two numerical methods, Newton and Gummel (Software 2023), have been considered to obtain better convergence when obtaining numerical solutions. Due to the channel's radius being more significant than 4 nm, the simulation does not consider the Quantum confinement effect (Tables 2, 3).

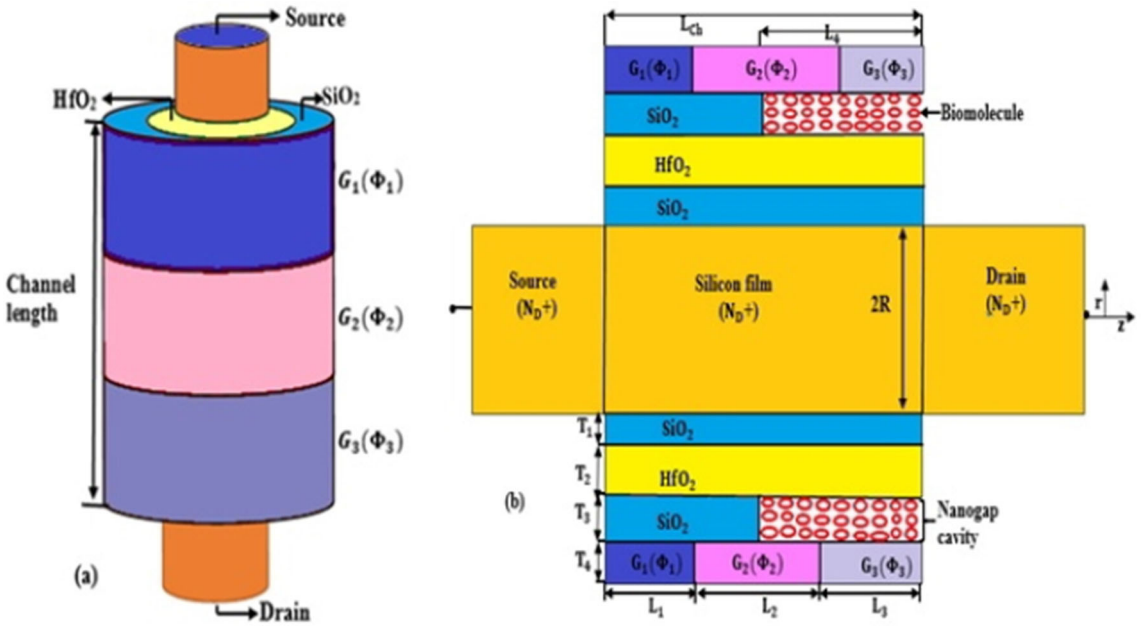


Fig. 3 Illustrates (**a**) 3D representation diagram and (**b**) 2D view of n-type triple hybrid metal gate dielectric modulated gate all around JL-NWFET

Table 1 Designed parameters for THM-high-k GAA-JL-SiNWFET

Parameters	Parameter values		
Channel length (nm)	20.00		
High-k (HfO_2) oxide thickness (nm)	1.50		
Interface oxide thickness SiO_2 (nm)	0.30		
Nanogap cavity length (nm)	10.00		
Length/thickness of source/drain (nm)	10.00		
Thickness of nanogap cavity (nm)	1.00		
Silicon film diameter (nm)	10.00		
Source/drain and channel doping (N_D +)	$10^{19}cm^{-3}$		
Oxide dielectric constant, HfO_2 and SiO_2)	25.00 and 3.90		
Gate work-function variables	Gate work-function values (Φ)		
	Φ_1(eV)	Φ_2(eV)	Φ_3(eV)
M	4.86	4.96	4.50
O	4.96	4.86	4.50
Q	4.86	4.50	4.96

2 Potential and subthreshold current modeling

2.1 Electrostatic potential analysis

A higher source to drain metal gate work function in small channel radius junctionless transistors enables the metal gate to deplete the silicon channel safely (Wang et al. 2011). The channel portion of the proposed device shown in Fig. 3(b) is divided into three distinct regions for the potential distribution's analytical expressions. Assuming uniform impurity doping concentration $\left(N_D = 1 \times 10^{19}cm^{-3}\right)$ through the channel. The $Si - SiO_2$ interface is considered to be free of trap charges. This model describes the movement of carriers through semiconductor material under the influence of electric fields (Henry 2007). Poisson's equation describes the electrostatic potential distribution inside the SiNWFETs and the surrounding medium. Poisson's equation relates the electric potential and the device's charge distribution. This case clarifies how the biosensor's electrostatics are impacted by the presence of biomolecules (uricase and ChO_X). In semiconductor devices, Poisson's equation often has the following form (Usha et al. 2020):

$$\nabla^2\boldsymbol{\phi} = \boldsymbol{\rho}/\boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ is the permittivity of SiNW, $\boldsymbol{\rho}$ is the charge density, and $\boldsymbol{\phi}$ the electrostatic potential.

Based on this assumption, Poisson's equation in corresponding regions ignoring source/drain depletion regions in the electrostatic potential $\varphi(r, z)$ calculation is written as;

Table 2 Model parameters applied in the simulations (Software 2023)

Physical models	parameter	Value and unit
(SRH) Shockley–Read–Hall carrier recombination	$TAUN_0$	1.0×10^{-7}S
	$TAUP_0$	1.0×10^{-7}S
(BEN) Bandgap narrowing	BGN.E	6.92×10^{-3}V
	BGN.N	$1.310^{17}cm^{-3}$
Lombardi CVT mobility	BN.CVT	$4.75 \times 10^{7}cm/(s)$
	BP.CVT	$9.925 \times 10^{6}cm/(s)$

Table 3 Constant parameters applied in this simulation (Software 2023)

Parameters	Symbols	Values and units
Electron affinity of silicon	χ_{Si}	~4.17eV
Silicon bandgap energy	E_g	~1.08eV
Permittivity of silicon	ε_{Si}	11.8
Boltzmann constant	K_B	$\sim 1.381 \times 10^{-23}$ J K^{-1}
Electronic charge	q	$\sim 1.6x10^{-19}$C
Room temperature	T	300 K
Planck's constant	h	$\sim 6.63 \times 10^{-34}$ JS
Silicon intrinsic carrier concentration	n_i	$\sim 1 \times 10^{10}cm^{-3}$

$$\frac{1}{r}\frac{\partial}{\partial r}\varphi_1(r,z)+\frac{\partial^2}{\partial r^2}\varphi_1(r,z)+\frac{\partial^2}{\partial z^2}\varphi_1(r,z)=-\frac{-qN_{D_1}}{\varepsilon_{Si}} \quad For \quad 0\le z\le L_1, 0\le r\le R \tag{1}$$

$$\frac{1}{r}\frac{\partial}{\partial r}\varphi_2(r,z)+\frac{\partial^2}{\partial r^2}\varphi_2(r,z)+\frac{\partial^2}{\partial z^2}\varphi_2(r,z)=-\frac{-qN_{D_2}}{\varepsilon_{Si}} \quad For \quad L_1\le z\le L_1+L_2, 0\le r\le R \tag{2}$$

$$\frac{1}{r}\frac{\partial}{\partial r}\varphi_3(r,z)+\frac{\partial^2}{\partial r^2}\varphi_3(r,z)+\frac{\partial^2}{\partial z^2}\varphi_3(r,z)=-\frac{-qN_{D_3}}{\varepsilon_{Si}} \quad For \quad L_1+L_2\le z\le L, 0\le r\le R \tag{3}$$

where L_1 L_2 and L_3 are region-I, region-II, and region-III channel/gate lengths. The channel's potential distribution is denoted by $\varphi_i(r,z)$"q" is the electronic charge, the homogeneous silicon film (channel) doping density is N_D, and the permittivity of silicon channel (Si) is ε_{Si}. The electrostatic potential $\varphi_i(r,z)$ is solved using the superposition technique and split into $V_i(r)$ and $U_i(r,z)$ (Software 2023) was written as:

$$\varphi_i(r,z)=V_i(r)+U_i(r,z) \quad for\ quad\ i=1,2 \text{ and } 3 \tag{4}$$

Applying a one-dimensional solution of Poisson's equation and a two-dimensional solution to the homogeneous Laplace equation with boundary conditions, for Eq. (4) i.e.

$$\frac{1}{r}\frac{\partial V_i(r)}{\partial r}+\frac{\partial^2 V_i(r)}{\partial r^2}=-\frac{-qN_D}{\varepsilon_{Si}} \quad \text{and } i=1,2 \text{ and } 3. \tag{5}$$

$$\frac{1}{r}\frac{\partial U_i(r,z)}{\partial r}+\frac{\partial^2 U_i(r,z)}{\partial r^2}+\frac{\partial U_i(r,z)}{\partial z}=0 \quad \text{and } i=1,2 \text{ and } 3. \tag{6}$$

Since the electric field at the center of the silicon channel has vanished (Profile 2018).

$$\left.\frac{\partial\varphi_i(r,z)}{\partial r}\right|_{r=0}=0,\ i=1,2 \text{ and } 3. \tag{7}$$

and there is a continuous electric field at the interface of (Si/SiO_2);

$$\left.\frac{\partial\varphi_i(r,z)}{\partial r}\right|_{r=R}=\frac{\varepsilon_{OX}}{\varepsilon_{Si}t_{OXeff}}\left[V_{gs}-\phi_{M_iS}-\varphi_i(r=R,z)\right], i=1,2 \text{ and } 3. \tag{8}$$

Here ϵ_{OX} is the dielectric constant of SiO_2, ϵ_{HfO_2} is the dielectric constant of HfO_2, t_{OXeff} is effective oxide thickness (EOT) obtained using the gate stack (SiO_2+HfO_2) (Ilatikhameneh et al. 2015; Garcia Ruiz et al. 2711) is given by;

$$t_{OXeff}=t_1+\frac{\varepsilon_{SiO_2}}{\varepsilon_{HfO_2}}t_2 \tag{9}$$

where t_1 and t_2 are the thickness of SiO_2 and HfO_2 respectively and Φ_{M_iS} is gate and channel work-function difference is given by;

$$\phi_{MS}=\phi_M-\phi_{Si} \tag{10}$$

where Φ_M stands metal work functions and ϕ_{Si} stands silicon channel work-function, written as (Software 2023),

$$\phi_{Si}=\chi_{Si}(T=300K)+q\phi_F+\frac{E_g(T,N_D)}{2q} \tag{11}$$

$$\phi_F=\frac{KT}{q}\ln\left(\frac{n_i}{N_D}\right) \tag{12}$$

Here, ϕ_F the channel's Fermi potential, χ_{Si} is the silicon electron affinity, n_i is inherent carrier concentration, and $E_g(T,N_D)$ is temperature and doping dependent on silicon bandgap energy (Software 2023),

$$E_g(T,N_D)=E_g(300k)+E_{g\alpha}\left(\frac{(300K)^2}{300K+Eg\beta}-\frac{T^2}{T+E_{g\beta}}\right)-\Delta E_g(N_D) \tag{13}$$

$E_g(300\text{ K})$ specifies the energy band gap at 300 K, $E_{g\alpha}=4.73\times10^4$eV/K and $E_{g\beta}=636$ K. $\Delta E_g(N_D)$ is the amount of energy needed to narrow the bandgap due to high doping (Software 2023); an analytic equation relates the variation in bandgap (ΔEg) due to doping concentration (N_D), using Atlas TCAD simulation tool as described in (Eq. 14) (Software 2023).

$$\Delta E_g(N_D)=\beta_E x\left\{\ln\left(\frac{N_D}{\beta_N}\right)+\left[\left(\ln\left(\frac{N_D}{\beta_N}\right)\right)^2+\beta_c\right]^{1/2}\right\} \tag{14}$$

where $\beta_E=6.92\text{meV}$, $\beta_N=1.3\times10^{17}\text{cm}^{-3}$ and $\beta_C=0.5$ (Software 2023).

From Eqs. (5)–(10) the solutions $V_i(r,)$ can be obtained as follows;

$$V_i(r)=\frac{qN_D}{4\varepsilon_{Si}}r^2+V_{gs}-V_{fbi}+\frac{qN_DR^2}{4\varepsilon_{Si}}+\frac{qN_DR}{2C_{OX}},\ i=1,2,3. \tag{15}$$

Here C_{OX} is the gate-oxide capacitance. Electron affinity and flat-band voltage of the channel is given as follows;

$$\chi_{Si(T)}=\chi_{Si(300K)}+\beta_c\left[E_{g(T)}-E_{g(300K)}\right]$$

$$V_{fbi} = \left\{ \phi_{M1} - \left(\chi_{Si(T)} + \frac{E_g(T, N_D)}{2} - q\phi_f(T) \right) \right.$$
$$For\ \mathrm{i} = 1,\ 0 \le \mathrm{z} \le \mathrm{L}_1$$

$$V_{fbi} = \left\{ \phi_{M2} - \left(\chi_{Si(T)} + \frac{E_g(T, N_D)}{2} - q\phi_f(T) \right) \right. \pm \frac{qN_F}{C_{OX}}$$
$$For\ \mathrm{i} = 2,\ \mathrm{L}_1 \le \mathrm{z} \le \mathrm{L}_1 + L_2$$

$$V_{fbi} = \left\{ \phi_{M3} - \left(\chi_{Si(T)} + \frac{E_g(T, N_D)}{2} - q\phi_f(T) \right) \right. \pm \frac{qN_F}{C_{OX}}$$
$$For\ \mathrm{i} = 3,\ \mathrm{L}_1 + L_2 \le \mathrm{z} \le \mathrm{L}_1 + L_2 + L_3 = L$$

The general solution for $\mathrm{U_i(r, z)}$ using the Fourier–Bessel series and separation of variable approach is (Profile 2018);

$$U_i(r,z) = \sum_{n=1}^{\infty} \left[A_{ni} e^{\left(\frac{\alpha_n z}{R}\right)} + B_{ni} e^{\left(\frac{-\alpha_n z}{R}\right)} \right] J_0\left(\frac{\alpha_n r}{R}\right), i = 1,2,3. \tag{16}$$

where $\mathrm{A_{ni}}$ and $\mathrm{B_{ni}}$ are coefficients and α_n is eigenvalue calculated using the following equation (Tolstov 1962) given as;

$$\frac{\varepsilon_{OX} R}{t_{OXeff}} J_0(\alpha_n) - J_1(\alpha_n)\alpha_n = 0 \tag{17}$$

The first kind of Fourier–Bessel function of order i is denoted by $\mathrm{J_i(x)}$ (Tolstov 1962). Coefficients of the Fourier–Bessel series $\mathrm{A_{n1}, B_{n1}, A_{n2}, B_{n2}, A_{n3},}$ and $\mathrm{B_{n3}}$ are determined based on the boundary conditions. Assuming that source-drain depleted region potential is independent of ″r.″ Eq. (1) is written using Poisson's equation;

(i) At the source end, electrostatic potential is written as;

$$\varphi_1(r,z)|_{z=0} = V_{bi} \tag{18}$$

Here $\mathrm{V_{bi}}$ is built potential.

(ii) At the drain end, electrostatic potential is written as;

$$\varphi_3(r,z)|_{Z=L} = V_{bi} + V_{ds} \tag{19}$$

where $\mathrm{V_{DS}}$ is drain bias voltage, and interface electric potential in the three regions is continuing and expressed as

$$\varphi_1(r,z)|_{Z=L_1} = \varphi_2(r,z)|_{Z=L_1} \tag{20}$$

$$\varphi_2(r,z)|_{Z=L_1+L_2} = \varphi_3(r,z)|_{Z=L_1+L_2} \tag{21}$$

Interface electric fields in the three areas are written as;

$$\left.\frac{\partial \varphi_1(r,z)}{\partial z}\right|_{Z=L_1} = \left.\frac{\partial \varphi_2(r,z)}{\partial z}\right|_{Z=L_1} \tag{22}$$

$$\left.\frac{\partial \varphi_2(r,z)}{\partial z}\right|_{Z=L_1+L_2} = \left.\frac{\partial \varphi_3(r,z)}{\partial z}\right|_{Z=L_1+L_2} \tag{23}$$

The coefficients $\mathrm{A_{n1}, B_{n1}, A_{n2}, B_{n2}, A_{n3},}$ and $\mathrm{B_{n3}}$ in Eq. (16) are obtained using Fourier–Bessel series and boundary conditions Eqs. (18–23) as (Tolstov 1962);

$$A_{n1} = \left[-T_{n1} e^{\left(\frac{-\alpha_n L}{R}\right)} + T_{n2} \cosh \frac{a_n(L_1 - L)}{R} + T_{n3} \cosh \frac{a_n(L_1 + L_2 - L)}{R} + T_{n4} \right] x \left(2 \sinh \frac{\alpha_n L}{R} \right)^{-1} \tag{24}$$

$$B_{n1} = \left[T_{n1} e^{\left(\frac{\alpha_n L}{R}\right)} - T_{n2} \cosh \frac{a_n(L_1 - L)}{R} - T_{n3} \cosh \frac{a_n(L_1 + L_2 - L)}{R} - T_{n4} \right] x \left(2 \sinh \frac{\alpha_n L}{R} \right)^{-1} \tag{25}$$

$$A_{n2} = \left\{ \left[-T_{n1} + T_{n2} \cosh \frac{\alpha_n L_1}{R} \right] e^{\left(-\frac{\alpha_n L}{R}\right)} - T_{n3} \cosh \frac{\alpha_n(L_1 + L_2 - L)}{R} + T_{n4} \right\} x \left(2 \sinh \frac{\alpha_n L}{R} \right)^{-1} \tag{26}$$

$$B_{n2} = \left\{ \left[T_{n1} - T_{n2} \cosh \frac{\alpha_n L_1}{R} \right] e^{\left(\frac{\alpha_n L}{R}\right)} + T_{n3} \cosh \frac{\alpha_n(L_1 + L_2 - L)}{R} - T_{n4} \right\} x \left(2 \sinh \frac{\alpha_n L}{R} \right)^{-1} \tag{27}$$

$$A_{n3} = \left\{ \left[-T_{n1} + T_{n2} \cosh \frac{\alpha_n L_1}{R} + T_{n3} \cosh \frac{\alpha_n(L_1 + L_2)}{R} \right] e^{\left(\frac{-\alpha_n L}{R}\right)} + T_{n4} \right\} x \left(2 \sinh \frac{\alpha_n L}{R} \right)^{-1} \tag{28}$$

$$B_{n3} = \left\{ \left[T_{n1} - T_{n2} \cosh \frac{\alpha_n L_1}{R} - T_{n3} \cosh \frac{\alpha_n(L_1 + L_2)}{R} \right] e^{\left(\frac{\alpha_n L}{R}\right)} - T_{n4} \right\} x \left(2 \sinh \frac{\alpha_n L}{R} \right)^{-1} \tag{29}$$

$$Where\ \mathrm{T_{n1}} = V_{bi} S_{no} - S_{n1} \tag{30}$$

$$T_{n2} = S_{n2} - S_{n1} \tag{31}$$

$$T_{n3} = S_{n3} - S_{n2} \tag{32}$$

$$T_{n4} = (V_{bi} - V_{DS}) S_{no} - S_{n4} \tag{33}$$

$$S_{no} = \frac{2J_1(\alpha_n)}{\alpha_n \left[J_1^2(\alpha_n) J_0^2(\alpha_n) \right]} \tag{34}$$

$$and\ \mathrm{S_{ni}} = V_{GS} S_{no} - U_{ni} \tag{35}$$

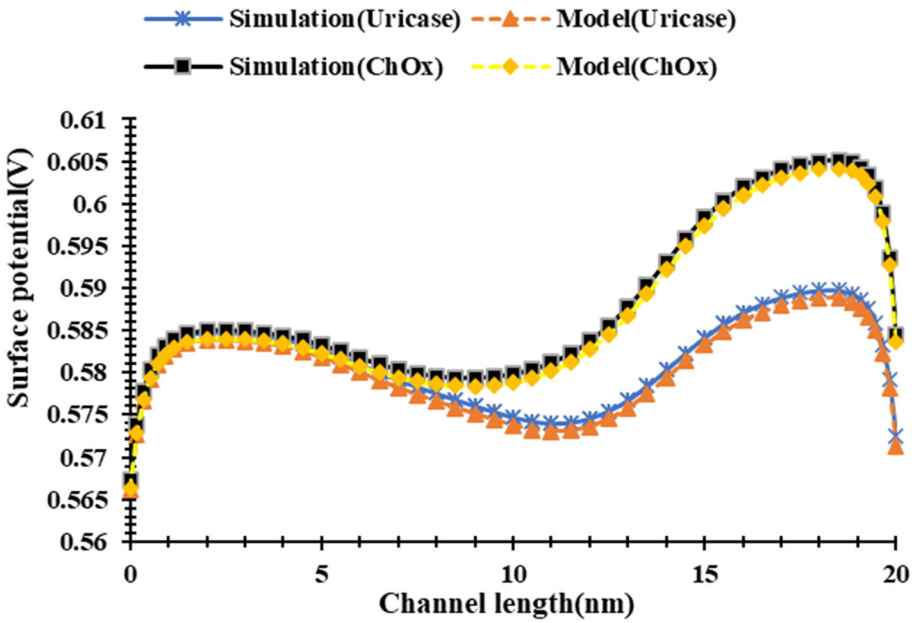


Fig. 4 Center potential along the channel for n-type triple hybrid metal gate dielectric modulated gate all around junctionless NWFET

$$where\ \mathrm{U}_{ni} = -\left[\Phi_{MiS} + \frac{qN_{Dit_{OXeff}}R}{2\varepsilon_{OX}} + \frac{qN_{Di}R^2}{4\varepsilon_{Si}}\right]S_{no} + \frac{qN_{Di}[R^2\alpha_n J_1(\alpha_n) - 2R^2 J_2(\alpha_n)]}{2\varepsilon_{Si}\left[J_1^2(\alpha_n) + J_0^2(\alpha_n)\alpha_n^2\right]} \quad (36)$$

Differentiating surface potential $\varphi(r = R, Z)$ with respect to z, gives channel electric field E(z) in the z-direction (Fig. 4).

$$E_i(z) = \sum_{n=1}^{\infty} A_{ni}e^{\left(\frac{\alpha_n z}{R}\right)} - B_{ni}e^{\left(\frac{\alpha_n z}{R}\right)} \times \left(\frac{\alpha_n}{R}\right)J_o(\alpha_n),\ \text{where i = 1,2,3} \quad (37)$$

2.2 Subthreshold current modeling

The center potential solution is used to analyze the subthreshold current (Bagga and Dasgupta 2017). In constant electron quasi-Fermi potential $\phi_i(z)$, carrier mobility from source to drain in the z-direction is determined using current density $J_i(z)$;

$$J_i(r,z) = -q\mu_n(r,t)\frac{d\phi_n(z)}{dz} \quad (38)$$

Here n(r, z) is inherent carrier concentration and μ_n is carrier mobility. Integrating the current density $J_i(r, z)$ along "z" in polar coordinates under a completely depleted channel length gives a drain current below threshold voltage (Localized 2011).

$$I_{Sub} = q\mu_n d\phi_i(z) \times 2\pi \int_0^R rn_i exp\{q[\phi(r,z) - \phi_i(z)]\}/KTdr \quad (39)$$

where "T" is thermal temperature, and "K" is Boltzmann's constant.

From z = 0 to z = L, I_{sub} expressed as;

$$I_{Sub} = KT\mu_n\left(n_i^2/N_a\right)\left\{1 - e^{\left(\frac{-q}{KT}V_{DS}\right)}\right\}/\int_0^L \left(1/2\pi \int_0^R re^{\frac{q\phi(r,z)}{KT}}dr\right)dz \quad (40)$$

where μ_n is silicon's constant mobility model ($\mu_n = 1417\mathrm{cm}^2/\mathrm{V.S}$).

2.3 *Threshold voltage (V_{th}) modeling*

The lowest (V_{GS}) wanted to establish a conducting path between the source and drain terminals in a MOSFET device is known as the threshold voltage (V_{th}) (Jean and Wu 1997). The threshold voltage for non-zero substrate bias voltage (V_{GS}) is written as;

$$V_{th} = \Phi_{MS} - 2\phi_F - 2\frac{Q_{BO}}{C_{OX}} - \frac{Q_B}{C_{OX}}$$
$$Where\quad Q_{BO} = -\sqrt{2qN_D\varepsilon_{si}|2\phi_F|},$$
$$Q_B = -\sqrt{2qN_D\varepsilon_{si}|-2\phi_F + V_{GS}|} \quad (41)$$

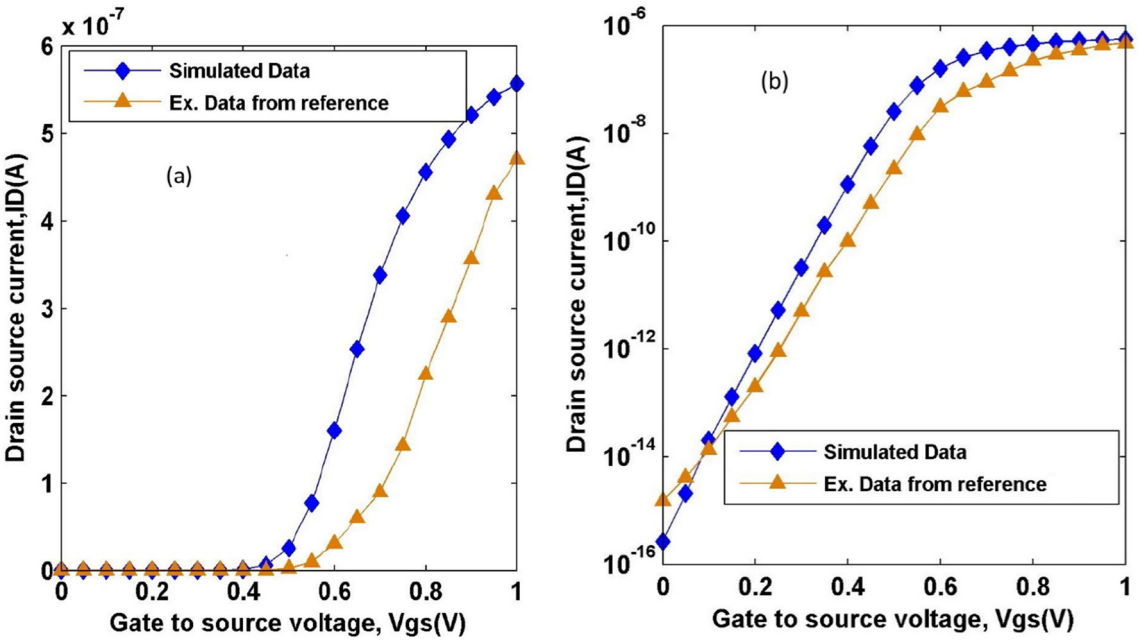


Fig. 5 Illustrates the simulation and experimental results (Im et al. 2007) of drain current for dielectric modulated triple hybrid metal gate all around JL-NWFET (**a**) IV in linear file and (**b**) IV in log file at (V_{DS} = 0.05 V)

where C_{OX} is gate-to-oxide capacitance, and Q_{OX} is gate-to-oxide charge density due to lattice imperfection at the interface.

2.4 Device calibration

In order to authenticate our work, the simulated results are in close proximity to the experimental results of Im et al. (2007) previously published works. Figure 5) illustrates the comparison of proposed device simulation output results with existing experimental data of references (Im et al. 2007) to validate our calibration. Figure 3a gives the ID-Vgs results in the linear scale, and Fig. 5b shows the ID–Vgs results in the log scale.

3 Results and discussion

3.1 Transfer characteristics

Device on-current and off-current are affected by increased dielectric constant biomolecule with different metal gate work-function, as shown in Fig. 6. The value of leakage

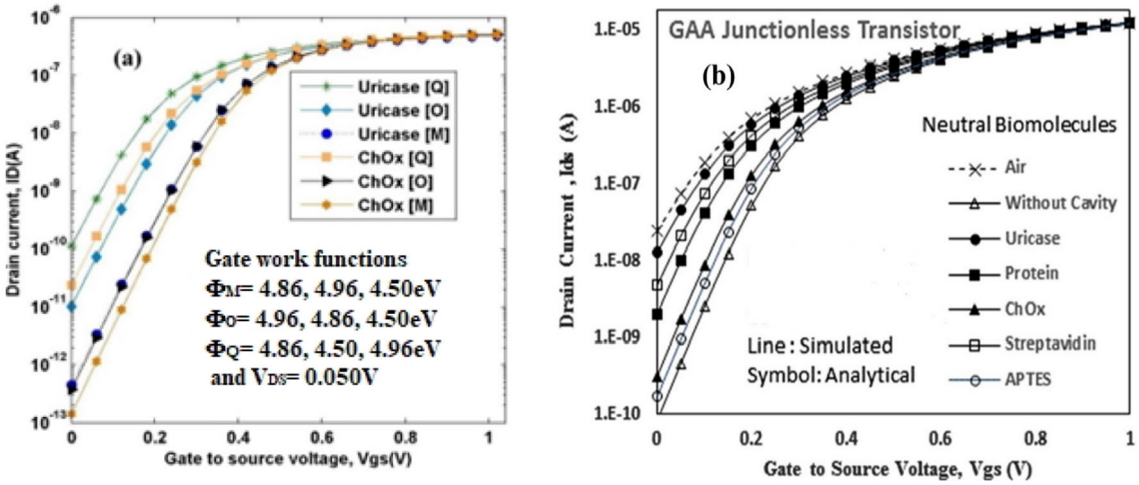


Fig. 6 Effect of different gate electrode work-function and dielectric materials on I_D-V_G and (**a**) proposed device and (**b**) referenced work (Pratap et al. 2018) biosensor devices

current indicated by M and Q is higher than that of "O." This indicates that the high gate electrode work-function sited by "O" suppresses tunneling current at weak inversion region enhance on current, as shown in Fig. 9b. For instance, the percentage variation of leakage current in all cases is 80.97% decreases when tunable gate work-function marked by "O" compared with "M" and 83.64% reduces when "O" is compared with "Q" using dielectric material ChO_X in the nanogap cavity. In the case of uricase dielectric material, the percentage variation is expressed as 59.90% decrease as a result of tunable gate work-function "O" is compared with "M" and 67.20% decrease when "O" is compared with that of "Q" at a constant drain to source current $V_{DS} = 0.05V$ and $V_{GS} = 1.0V$). We can elucidate that higher dielectric materials with tunable gate electrode work function signify CMOS devices' performance at ultra-small scale device dimensions.

3.2 Effect of gate electrode work function on threshold voltage (V_{th})

A critical technological challenge for the triple gate MOSFET device is to find gate electrodes with appropriate tunable electrode work functions designed for the desired (V_{th})/threshold voltages (Localized 2011). The work function difference between the gate electrode and the intrinsic silicon film can be adjusted to optimize the threshold voltages for n-channel devices in a CMOS circuit. (Lin et al. 2002). Figure 7a illustrates the variation of threshold voltage when gate electrode work function increases from drain to source (denoted by "O" leads to the lower tunneling path, thereby raising the tunneling probability and reducing the ambipolar conduction (both positive and negative ions can move in the opposite direction) (Bagga and Dasgupta 2017). Replacing a higher dielectric constant in the nanogap cavity increases threshold voltage, as described in Fig. 7a. The shift threshold voltage (V) is illustrated in Fig. 7b, which signifies the sensitivity of proposed devices with a tunable work function labeled at "O," which leads to improved device performance. Varying metals have varying work functions, resulting in changes in the energy barrier at the metal–semiconductor interface and influencing carrier injection efficiency and transistor properties like threshold voltage. The calculation of shifting the threshold voltage (ΔV_{th}) is described by Eq. (42) (Rahman et al. 2017).

$$\Delta V_{th} = V_{th(Bio)} - V_{th(Air)} \tag{42}$$

Figure 8a depicts the effect of various gate electrode work-function device sensitivity ($S_{I_{OFF}}$) in the presence of biomolecules. It signifies that as a biomolecule dielectric constant increases, $S_{I_{OFF}}$ increased at optimized gate work-function tunned at "O" compared with" M, and "Q." For instance, percentage change /improvement of current sensitivity ($S_{I_{OFF}}$) due to different gate work functions are 16.21% when "O" compared with "M" and 30.96% compared to "Q" at ChO_X dielectric material when $V_{DS} = 0.05V$ &$V_{GS} = 1.0V$. We examine that ChO_X dielectric material and gate work-function sited by "O" improve current sensitivity compared to uricase dielectric material. Current sensitivity of the proposed device for uricase and ChO_X dielectric material, including different gate work-function, is obtained using Eq. (43) (Pratap et al. 2018)

$$S_{I_{off}} = \left.\frac{I_{off}(with\ biomolecule)}{I_{off}(without\ biomolecule)}\right|_{V_{GS}=0V} \tag{43}$$

Not only sensitivity ((S_{Ioff}) of our device is improved over the referenced works, but also switching ratio, surface potential, short channel effects(SCEs), and leakage /tunneling currents are more enhanced compared to the referenced (Sze 2014) shown in Fig. 8b**,** this is due to new

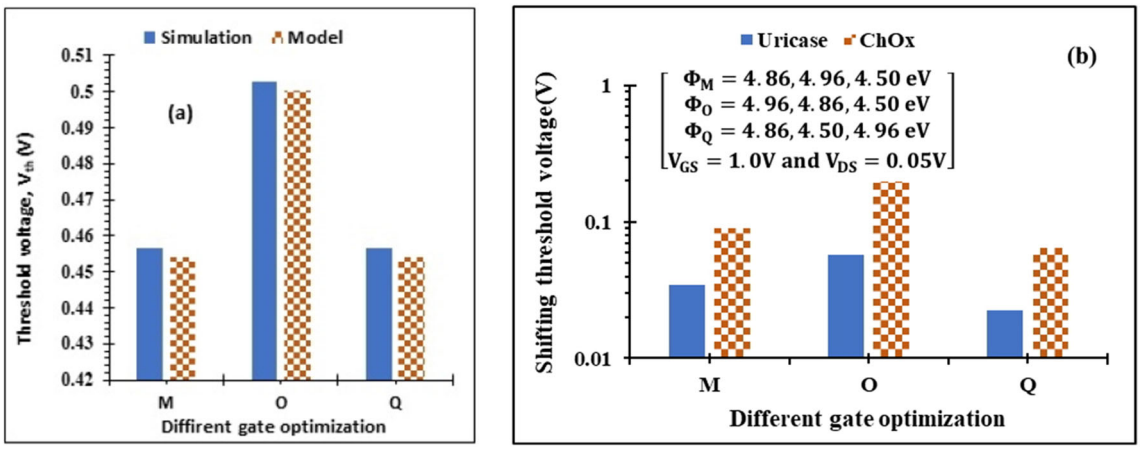


Fig. 7 Effect of different gate engineering on (**a**) threshold voltage and (**b**) shifting threshold voltage of n-type triple hybrid metal gate high-k gate all around junctionless NWFET

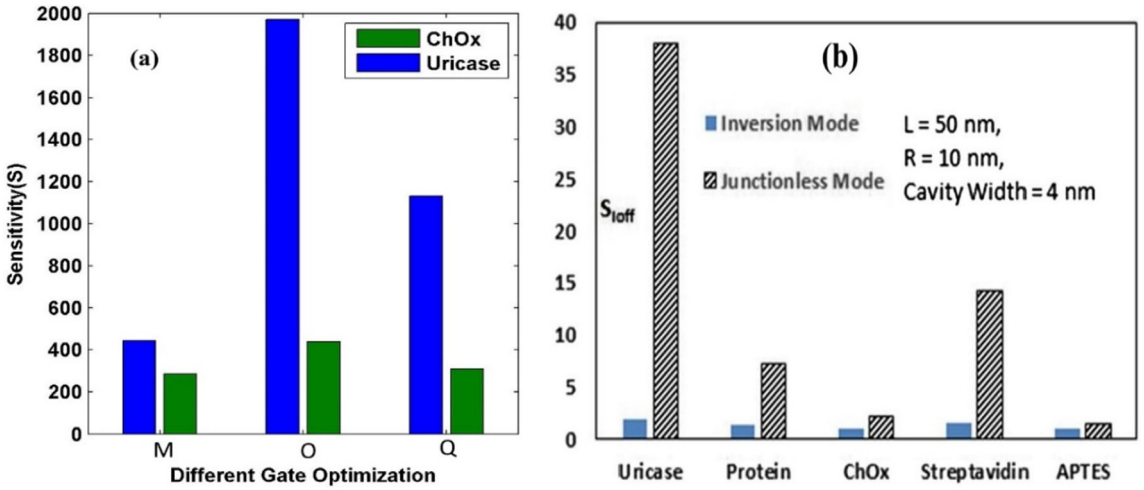


Fig. 8 Sensitivity parameter (S_{Ioff}) for (**a**) proposed and (**b**) past work (online published) devices (Pratap et al. 2018)

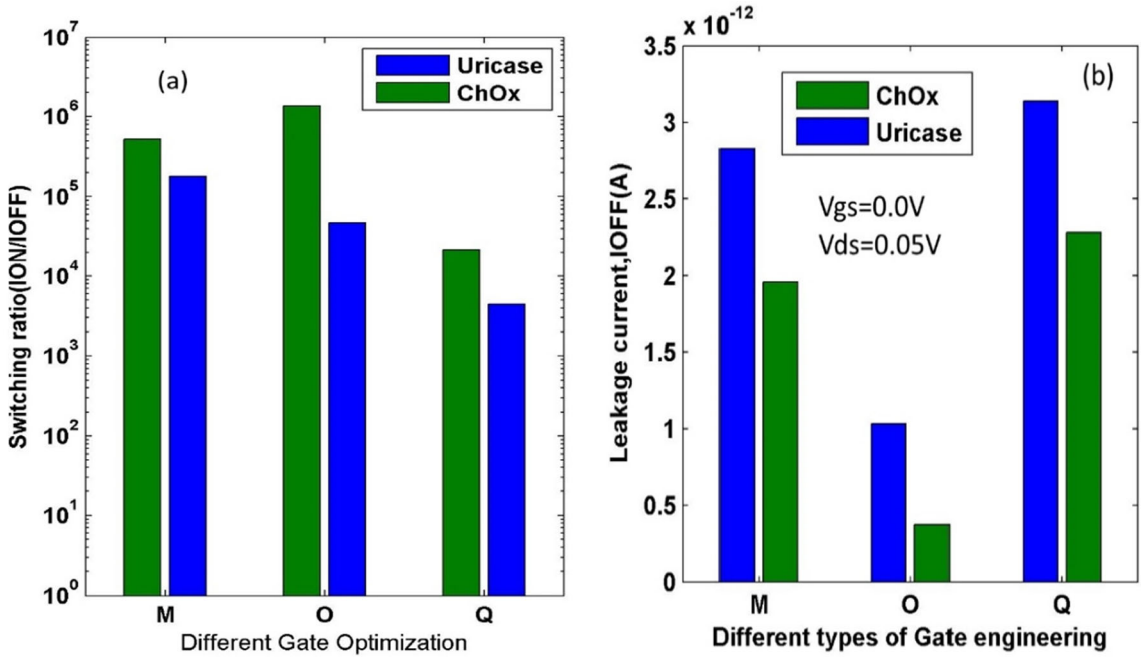


Fig. 9 Impact of different gate engineering on (**a**) switching ratio and (**b**) leakage current for n-type triple hybrid metal gate dielectric modulated GAA-JL-NWFET at $V_{DS} = 0.05V$ and $V_{GS} = 1.0V$, including uricase and ChO_X biomolecules

technologies incorporated in our device modeling, such as device dimension, device structure, high-k gate oxide material (HfO_2), and high optimized gate electrode work-function near to source side with higher electrode work function.

Figure 9 illustrates the switching ratio for different gate electrode work-function and leakage current of the proposed device. Increasing the on–off current ratio indicates overall improvement in drive current with a low leakage current in low stand power (LSTP) logic technology, even though the on-current decreased with a rise in gate electrode work-function near the source side. (James et al. 2010). Compared with M and Q gate tunable work function, the highest switching ratio is achieved at "O." Also, the device performance is enhanced at "O," as illustrated in Fig. 8a, due to higher electrode work function near the source side screens and drives carries across the channel (Bagga and Dasgupta 2017). The concentration of the target analytes close to the biosensor's sensing zone rises in the presence of uricase and ChO_X proteins. As a result, the electrical signal produced by the analytes' contact with the biosensor is amplified, improving sensitivity and lowering detection limits.

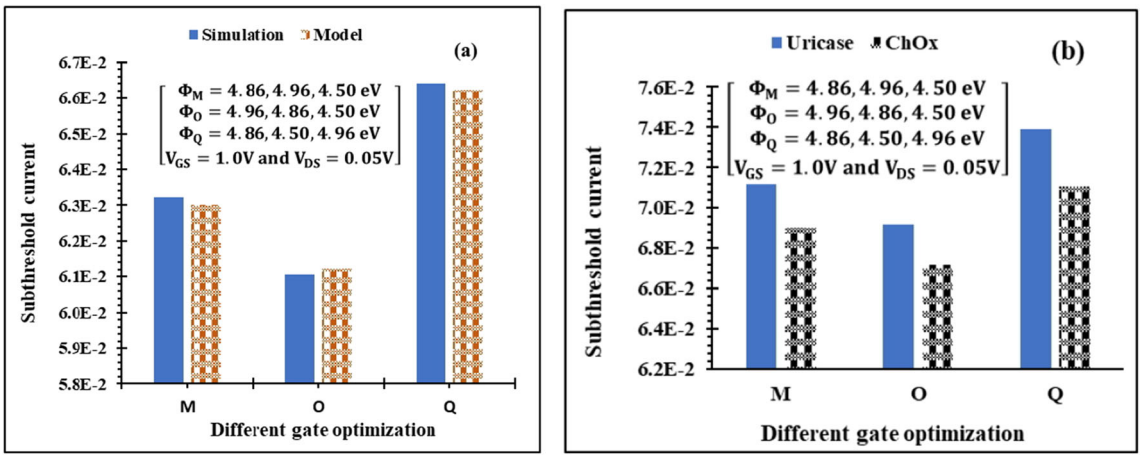


Fig. 10 Effect of different gate engineering on (**a**) analytical and simulation result of subthreshold slope and (**b**) variation of subthreshold slope along uricase and ChO_X biomolecules for n-type triple hybrid metal gate dielectric modulated junctionless gate all around NWFET

Figure 10 shows a lower sub-threshold slope when the gate has a tunable electrode work-function sited "O" while at 'M" and "Q," sub-threshold slope slightly increases due to the short channel effect at ultra-small channel length and hot carriers. Meanwhile, threshold voltage and drain mobility increase when the gate work function increases from source to drain through the channel designed at "O."

3.3 Transconductance (g_m) and output conductance (g_d)

Transconductance (g_m) measures drain current variation with a gate-source voltage (V_{GS}), keeping the drain-source voltage (V_{DS}) constant to describe device load drivability (Madan and Rishu 2016; Sze 2014; Van Zeghbroeck 2008). The switching speed of a circuit and voltage gain of MOSFET amplifiers are also characterized by transconductance. Since linearity and distortion are crucial issues in designing several circuits (Chaujar et al. 2009). The highest transconductance describes the circuit's capability for high-speed operation (Saib et al. 2015). The nonlinearity of the device can be scrutinized by determining the higher-order derivatives of $I_{DS} - V_{GS}$ at a constant drain-source voltage, and it designates the harmonic distortion. "Better linearity and lower distortion are examined under higher-order transconductance with lower amplitude" (Gupta and Chaujar 2015). The immobilized dielectric constant in the nanogap cavity affects device operation/linearity, as inferred by Fig. 11. Device transconductance (g_m) in terms of drain current and the gate-source bias voltage is given by Eq. (44) (Van Zeghbroeck 2008).

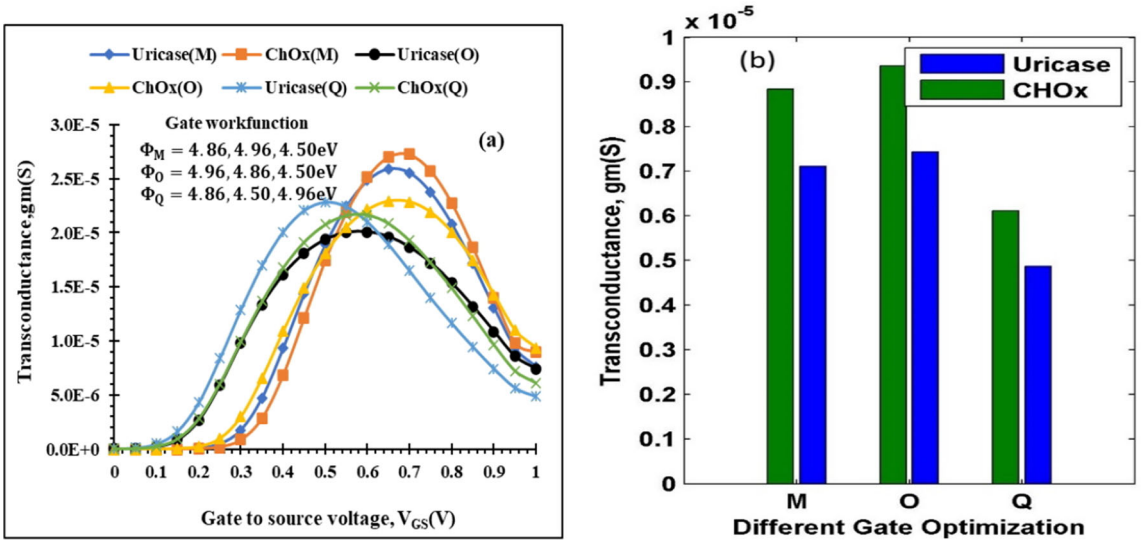


Fig. 11 Effect of different gate engineering on (**a**) transconductance and (**b**) transconductance at ($V_{DS} = 0.6V$and$V_{GS} = 1.0V$) for n-type triple hybrid metal gate dielectric modulated gate-all-around junctionless NWFET including uricase and ChO_X biomolecules

$$g_m = \left.\frac{\partial I_D}{\partial V_{GS}}\right|_{V_{DS}=0.6V} = \sqrt{\mu_n C_{OX}(\frac{W}{L})I_D} = \frac{2I_D}{V_{GS}-V_{th}} \quad (44)$$

$$g_d = \left.\frac{\partial I_D}{\partial V_{DS}}\right|_{V_{GS}=0.6V} = \left(\mu_n C_{OX}\frac{W}{L}\right)(V_{GS}-V_{th}-V_{DS}) \quad (45)$$

In the case of the gate-to-source voltage, variation of transconductance, including different gate electrode work-function, is illustrated in Fig. 11a. It clearly shows that, as gate work function increases from drain to source through the channel (labelled tunable work function at "O"), higher transconductance is obtained, as shown in Fig. 11b when V_{GS} is 1.0V and V_{DS} is 0.6V due to improved gate controllability as a result of high electrode work function near to source side**.** For instance, the percentage changes of transconductance labelled by "O" is 5.88% compared to "M" and 52.94% compared to "Q" using ChO_X dielectric material and 4.49% compared to "M" and 52.77% compared to "Q" using uricase dielectric material, labelled at "O." This suggests inadequate gate engineering work-function and dielectric materials also affects the CMOS device performance at ultra-small scale channel length.

Another critical parameter that indicates CMOS linearity is drain conductance, defined with different drain currents in varying drain voltage at constant gate voltage (Van Zeghbroeck 2008). Figure 12) illustrates the proposed device output conductance with different gate work functions noted by "M, O, and Q," including uricase and ChO_X dielectric materials itching on the nanogap cavity. In the lower inversion region, higher output conductance is experienced, while as bias voltage increases, the drain conductance decreases and comes to cloth each other, as illustrated by Fig. 12a. The output conductance (g_d) in terms of drain current and a drain-source bias voltage is given by Eq. (45) (van Zeghbroeck 2008).

The decreasing percentage change of output conductance labelled at "O" is 36.42% and 43.65% compared to labelled work functions of "M" and "Q," respectively, at ChO_X a dielectric material, while at uricase dielectric materials are 30.68% and 40.92%. This signified that gate engineering work function and coated dielectric materials of nanocavity affect the CMOS device's performance. As a result, the electrical signal produced by the analytes' contact with the biosensor is amplified, and the sensitivity is improved.

The transistor's intrinsic voltage gain ($g_{m/gd} or g_m R_O$) signifies the maximum voltage gain provided by a single transistor, playing a vital role in high-gain amplifiers. Figure 13. illustrates the variation of inherent voltage gain with different gate work-function, including dielectric gate material in the nanogap cavity noted by M, O, and Q. It is observed that at dielectric material, ($ChO_X, \epsilon = 3.3$) tunable work-function labelled by "O" is much higher compared to that of "M and Q." The degradation of the output resistance under work-function marked by "M and Q" of scaled CMOS technology is moderately caused by the DIBL effect (Sengupta and Pandit 2014). As illustrated in Fig. 13b, the intrinsic voltage gain is high at the "O" work function compared to "M and Q" at $V_{DS} = 0.6V$ and $V_{GS} = 1.0V$. The percentage change or improvement of inherent voltage gain, labelled by "O," is 31.58% compared to that of "M" and 187.61% compared to that of "Q" at ChO_X dielectric material. In the same manner, at uricase dielectric material, transconductance

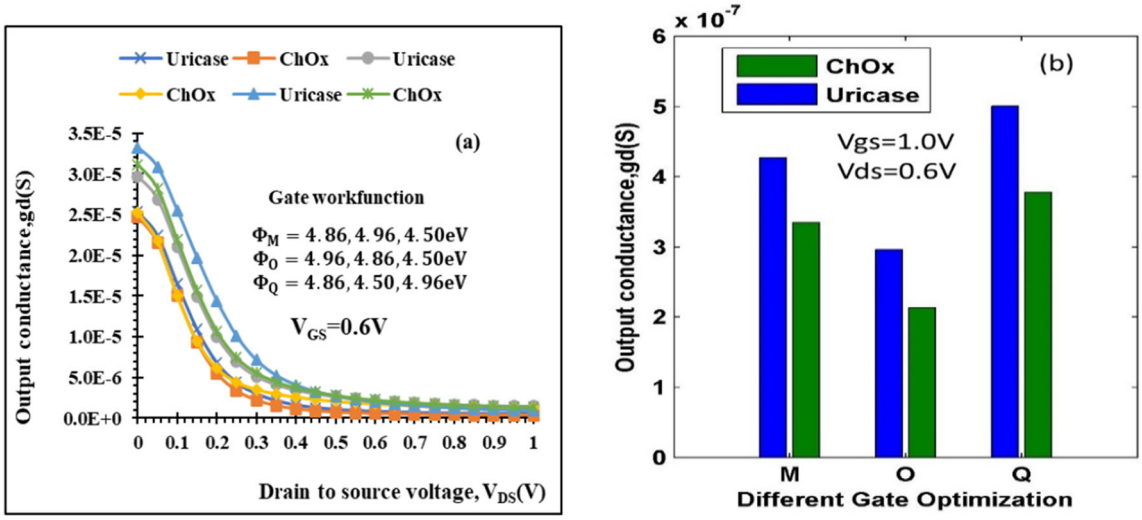


Fig. 12 Effect of different gate engineering on (**a**) output conductance and (**b**) output conductance for n-type triple hybrid metal gate dielectric modulated junctionless gate-all-around NWFET, including uricase and ChO_x biomolecules

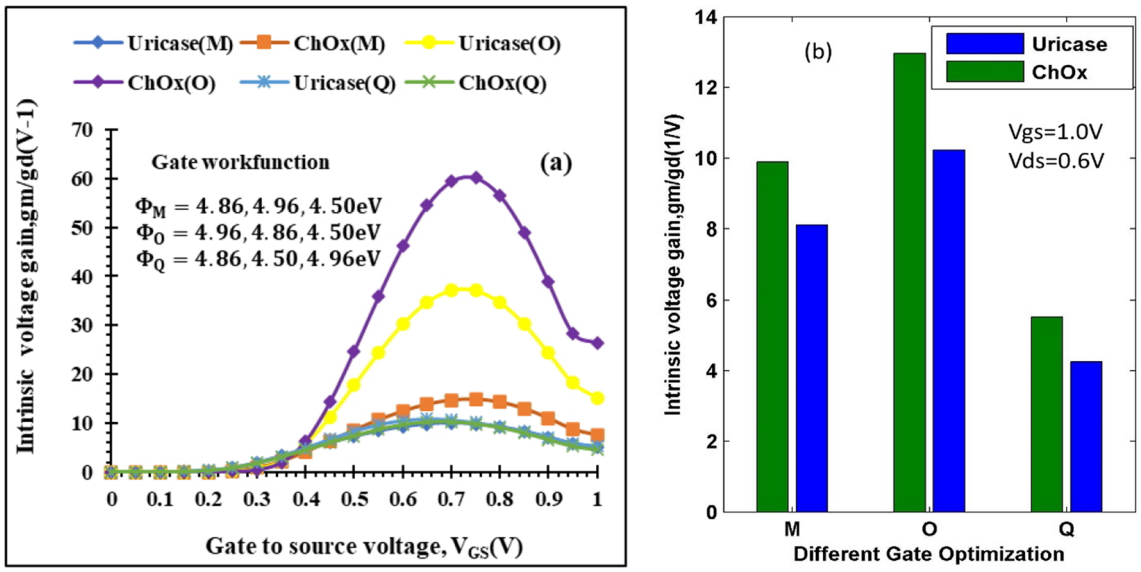


Fig. 13 Effect of different gate engineering on (**a**) intrinsic voltage gain ($g_m R_O$), (**b**) intrinsic voltage gain for n-type triple hybrid metal gate dielectric modulated gate-all-around junctionless NWFET including uricase and ChO_X biomolecules

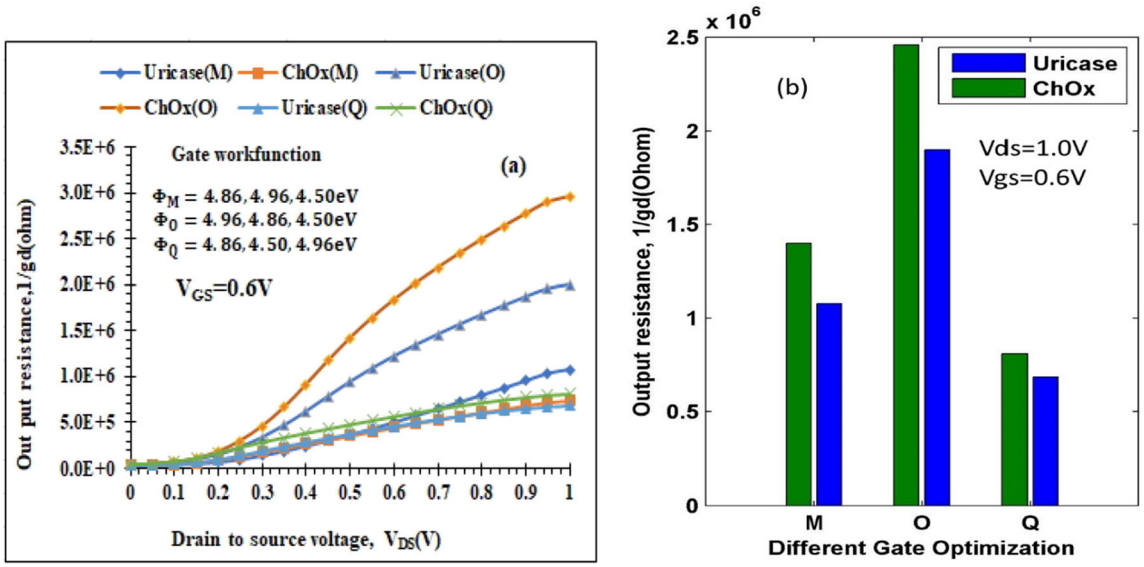


Fig. 14 Impact of different gate engineering on (**a**) output resistance and (**b**) output resistance for n-type triple hybrid metal gate dielectric modulated gate-all-around junctionless NWFET, including uricase and ChO_X biomolecules

labelled by "O" is 25.93% compared to "M" and 94.29% compared to "Q." This signifies gate engineering work-function with the nanogap cavity dielectric materials affects the intrinsic voltage gain and CMOS device performance. Mathematical expressions are written as (Sengupta and Pandit 2014).

$$\text{Intrinsic voltage gain (G)}\ \frac{g_m}{g_d} = g_m R_0 \tag{46}$$

where R_o is the output resistance given by $\frac{1}{g_d} =$ (Yuan and Woo 2005).

Figure 14 illustrates the effect of gate engineering labelled "M, O, and Q" on the output resistance. It is observed that the tunnelable work function marked by "O" has the highest output resistance than "M and Q." The gate engineering design at "O" experienced a less short channel effect due to increasing work function from drain to source, providing more mobility and gate controllability across the channel. Figure 14b clearly describes the effects of coated dielectric material on the output resistance. The percentage change of output resistance labelled by "O" is 75.71% compared to "M" and 203.70% compared to "Q" at ChO_X

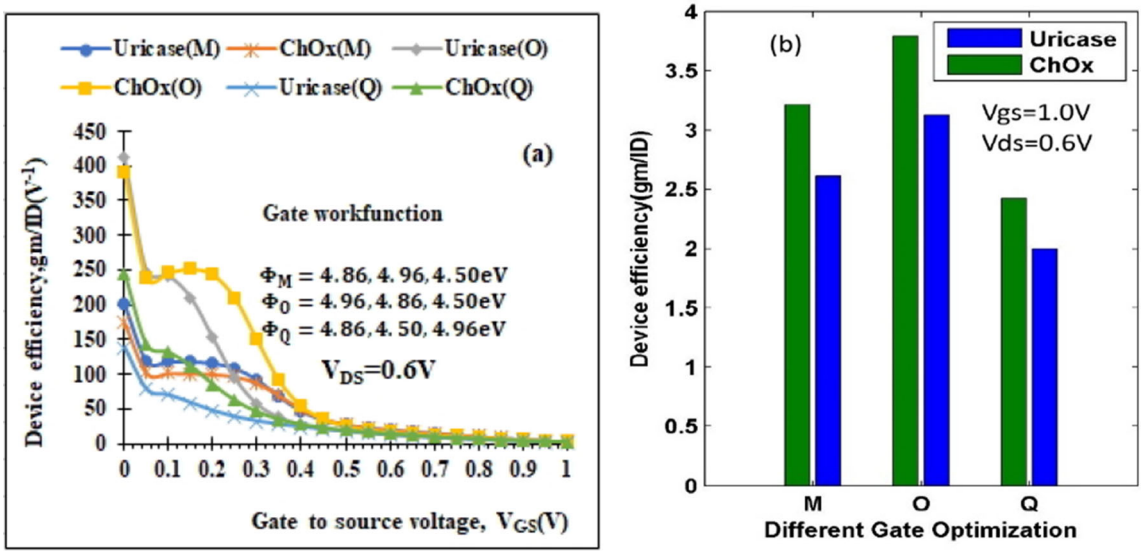


Fig. 15 Effect of different gate engineering on (**a**) device efficiency and (**b**) device efficiency for n-type triple hybrid metal gate dielectric modulated gate-all-around junctionless NWFET, including uricase and ChO_X biomolecules

a dielectric material. For uricase dielectric material, the increment of output resistance labelled by "O" is 75.93% compared to "M" and 167.97% compared to "Q." This signifies gate engineering work function, including nano-gap cavity dielectric material, affects the CMOS device performance.

Figure 15a illustrates the transconductance ratio over drain current (g_m/I_D), including different gate electrode work-function labelled by "M, O, and Q," undercoated cavity with uricase and ChO_X dielectric material signifies maximum value in weak inversion (WI) and decreases the value in strong inversion region. Figure 15b clearly describes the impact of uricase and ChO_X dielectric material when $V_{DS} = 0.6$ V and $V_{GS} = 1.0$ V over-device efficiency. Percentage increment of device efficiency (g_m/I_D) designated at "O" is 18.01% compared to "M" and 56.38% compared to "Q" at ChO_X dielectric material, and for Uricase dielectric material, the increment of device efficiency labelled by "O" is 19.92% compared to "M" and 56.5% compared to "Q." This improvement is due to potential barrier reduction in the channel due to higher electrode work-function and dielectric material near the source side enhances gate controllability, resulting in speed-to-power dissipation performance (Chaujar et al. 2009). This signifies gate engineering work function and dielectric materials in the nanogap affect CMOS device performance.

Figure 16 shows how cavity length impacts device transfer properties with various biomolecules. Simulated results were obtained utilizing cavity lengths of 5 nm,

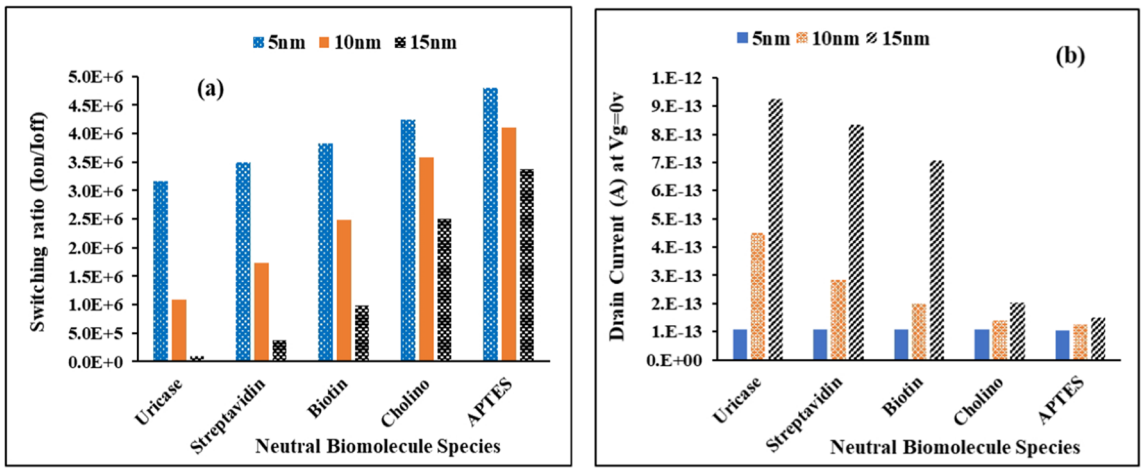


Fig. 16 Directs the impact of nanogap cavity length variation on (**a**) switching ratio and (**b**) leakage current for different biomolecules in triple hybrid metal gate dielectric modulated gate-all-around junctionless NWFET

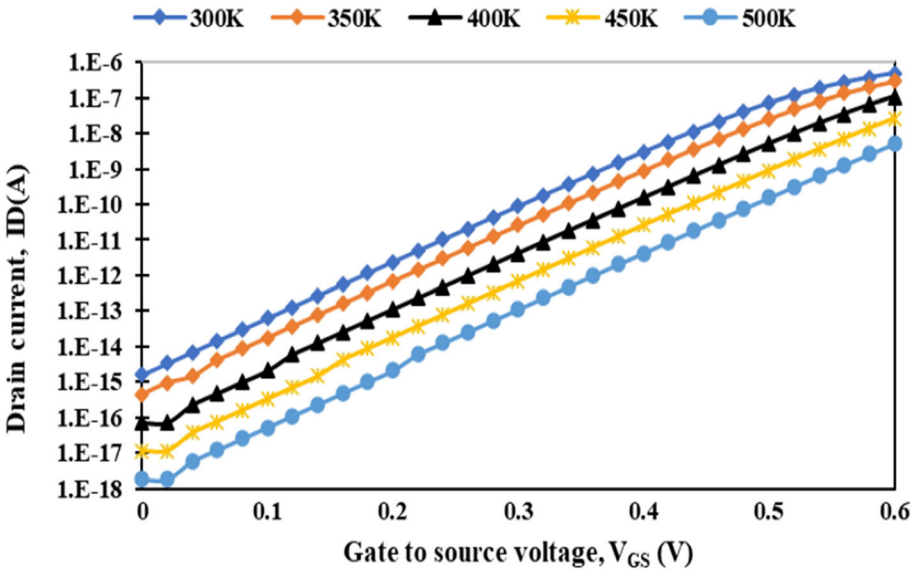


Fig. 17 Impact of temperature variations on I_D-V_{GS} characteristics of triple hybrid metal gate dielectric modulated gate-all-around junctionless NWFET

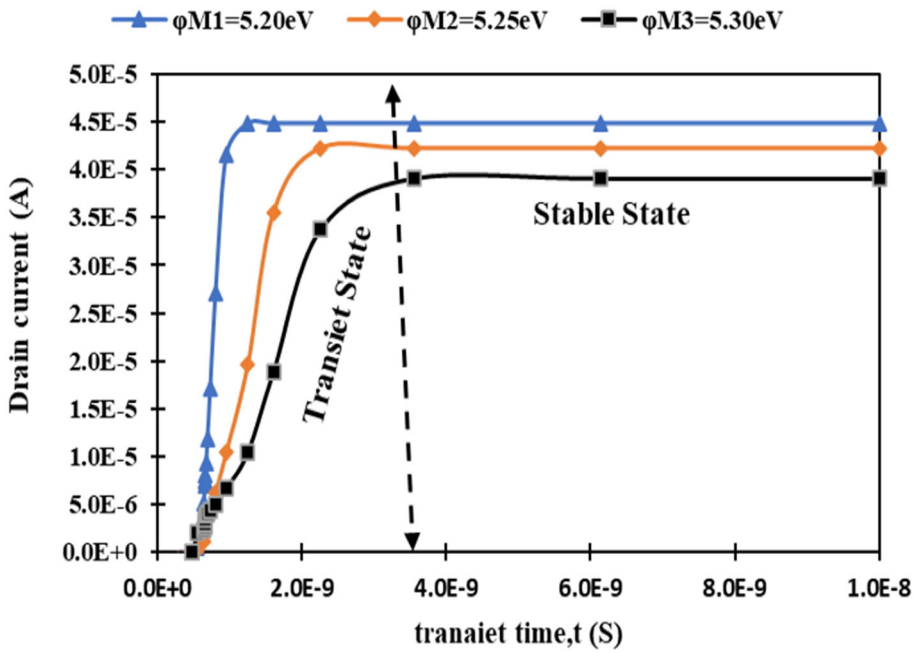


Fig. 18 Transient solution of the drain current in triple hybrid metal gate dielectric modulated gate-all-around junctionless NWFET at different gate electrode work functions

Table 4 Summarizes a comparison between the performance indicators for the two materials (uricase and ChOx) for n-type triple hybrid metal gate dielectric modulated gate-all-around junctionless NWFET at a constant gate to source voltage (V_{GS} = 1.0 V)

Gate work-function variables (eV)	Intrinsic voltage gain (V^{-1})		Switching ratio (Ion/Ioff ratio)		Output resistance ($1/g_d$)		Device efficiency		Transconductance, g_m (S)	
	Uricase	ChO_X	Uricase	ChO_X	Uricase	ChO_X	Uricase	ChO_X	Uricase	ChO_X
M	8.10	9.88	1.78×10^5	5.22×10^5	1.08×10^6	1.40×10^6	2.61	3.22	7.12×10^{-6}	8.84×10^{-6}
O	10.20	13.00	4.75×10^5	1.36×10^6	1.90×10^6	2.46×6	3.13	3.80	7.44×10^{-6}	9.36×10^{-6}
Q	4.25	5.22	4.40×10^3	2.18×10^4	6.85×10^5	8.10×10^5	2.00	2.43	4.83×10^{-6}	6.12×10^{-6}

10 nm, and 15 nm while maintaining gate length, doping profile, and nanogap cavity thickness, keeping constant at constant room temperature. Figure 16a shows how different biomolecules affect the switching ratio at various cavity lengths. Higher switching ratio variation is examined when nanogap cavity length decreases due to lower parasitic capacitances. It's also interesting to note that the nanoscale size gap significantly affects the electrical characteristics of FETs. As a result, a significant change in switching ratio (due to the dielectric constant and cavity length variation) is used as a detection/sensing metric for a specific biomolecule. Moreover, one of the critical parameters for the developed biosensor device is sensitivity. Due to the increase in nanogap cavity length, leakage current/tunneling current increases, as reflected in Fig. 16b, due to increasing tunneling distance (Tripathi et al. 2019)

Overall, we conclude that the cavity gap length must be optimized for better performance, allowing sub-20-nm THM-high-K-GAA-JL-NWFETs to be used as high-speed, low-noise, and high-sensitivity biosensors for the detection of a variety of diseases. The methodology utilized to investigate the impact of nanogap cavity length on the device performance most likely involves numerical simulations within the framework of the proposed device. The effect of charge confinement, electrostatic control, and biomolecule detection sensitivity was examined. Insights are gained into the appropriate. To assess the validity of the findings, the simulation results were compared with the theoretical models and current literature. Discrepancies or limitations in the model are identified, and recommendations for future development or refinement are made.

3.4 Long-term stability or reliability analysis

Physical characteristics, sensitivity, stability, and reproducibility are vital analyses for assuring the device's reliability over time and between devices (Matsukawa et al. 2008) at different temperatures. The temperature has various effects on the sensor system, and understanding these effects is critical for maintaining dependable and accurate sensor performance (Pour and Aval 1993). Consequently, the stability of our proposed sensor is further tested by investigating the influence of temperature on the transfer characteristics, as depicted in Fig. 17.

Also, analyzing a device's transient reaction can provide insights into the device's stability characteristics. A biosensor with a rapid and constant transient response would most likely be considered more stable. As a result, measuring the transient response of the biosensor optimized using numerical modeling for triple hybrid gate optimization can provide useful information about its stability under various operating settings, environmental influences, and during long periods of use. Figure 18 depicts the transient response of the drain current at different metal gate electrode work functions of the proposed device.

4 Conclusion

Based on a 2D solution of Poisson's equation, a new numerical model for triple hybrid metal gate dielectric modulated gate-all-around junctionless nanowire FET-based uricase and ChO_X biosensor was successfully developed to investigate the effects of gate electrode work-function during gate engineering/optimization. The numerical modeling technique simulates the electrical behavior of the biosensor using various combinations of metal work-function values. The selected metal work-function values are also crucial for gate engineering of SiNWFET-based biosensor devices, as they improve the device's sensitivity and performance in detecting biomolecules. Due to a systematic investigation and optimization of the hybrid gate structure along different gate electrodes, the theoretical modeling and simulation results are in good agreement and accurately predict the device's behavior. The analytical modeling contains an electrostatic potential, threshold voltage, subthreshold current, and electric field, including neutral biomolecules (uricase and ChO_X) and different metal work-function. The effect of gate electrode work function and dielectric materials on device sensitivity, linearity (transconductance, intrinsic voltage gain, output conductance), switching ratio, leakage current, subthreshold voltage, device efficiency, and shifting threshold voltage has been studied. The simulated results reveal that the voltage and current sensitivity increase when the tunable gate work function increases from drain to source across the channel. For instance, sensitivity improvement (S_{Cur}) is 16.21% when "O" compared with "M" and 30.96% compared to "Q" when ChO_X, a dielectric material, is immobilized on the cavity at designed gate engineering denoted by "O." This improvement of current and voltage sensitivities makes the THM-GAA-JL-NWFET based ChO_X biosensor designed at "O" clearly indicates better output results compared to "M" and "Q" in terms of sensitivity, switching ratio, and device efficiency, as it is summarized in Table 4.

In addition to the metal work function, higher dielectric constant of biomolecules and properly optimized nanogap cavity length signifies improved output results in transconductance, surface potential, switching ratio, and intrinsic voltage gain so that THM-GAA-JL-NMOSFET based ChO_X biosensor designed at "O" electrode optimization is responsible for better detection different concentration levels of cholesterol and uric acids. For instance,

hyperuricemia, gout, cardiovascular disease, hyperlipidemia, and metabolic syndrome are some specific diseases that could be detected using the proposed biosensor.

Acknowledgements The authors thank Ethiopia's Ministry of Higher Education and Microelectronics Research Laboratory, Delhi Technological University, Delhi, India, for providing all the necessary materials for our research.

Author contributions Rishu Chaujar implementation, review, editing at different stages, and analysis. Mekonnen Getnet Conceptualization, writing-original draft, software, methodology, analysis, and data curation.

Funding This research was not supported financially.

Availability of data and material The authors mentioned above have all relevant data related to this study effort and will be willing to disclose it if asked to do so in the future.

Declarations

Conflict of interest The authors declare they have no competing interests.

Compliance with ethical standards The authors have reviewed all of the Ethical Standards and are expected to adhere to them in the future.

Consent to participate and for the publication Since the study report is for a 'non-life science journal,' it is 'Not Applicable' in this case.

References

Ajay RN, Saxena M, Gupta M (2015) Investigation of dielectric modulated (DM) double gate (DG) junctionless MOSFETs for application as a biosensors. Superlattices Microstruct 85:557. https://doi.org/10.1016/j.spmi.2015.04.040

Ambhorkar P, Wang Z, Ko H, Lee S, Koo KI, Kim K, Cho DID (2018) Nanowire-based biosensors: from growth to applications. Micromachines 9:1. https://doi.org/10.3390/mi9120679

Ana F, Din NU (2011) Gate workfunction engineering for deep submicron MOSFET's: motivation, features and challenges. Int. J. Electron. Commun Technol. 2:29

Avila ES, Tinoco JC, Member S, Martinez-lopez AG, Reyes-barranca MA, Cerdeira A, Member S (2016) Parasitic gate resistance impact on triple-gate FinFET CMOS inverter. IEEE Trans Electron Devices 63:2635. https://doi.org/10.1109/TED.2016.2558580

Bagga N, Dasgupta S (2017) Surface potential and drain current analytical model of gate all around triple metal TFET, vol XX, p 1. https://doi.org/10.1109/TED.2016.2642165

Biswal SM, Swain SK, Sahoo JR, Swain AK, Routaray K, Nanda U, Biswal B (2019) A comparative study of junctionless triple-material cylindrical surrounding gate tunnel FET, vol 521. Springer, Singapore. https://doi.org/10.1007/978-981-13-1906-8_80

Chau R, Datta S, Doczy M, Doyle B, Kavalieros J, Metz M (2004) High-κ/metal-gate stack and its MOSFET characteristics. IEEE Electron Device Lett 25:408. https://doi.org/10.1109/LED.2004.828570

Chaujar R, Kaur R, Saxena M, Gupta M, Gupta RS (2008) Laterally amalgamated DUal material GAte Concave (L-DUMGAC) MOSFET for ULSI. Microelectron Eng 85:566. https://doi.org/10.1016/j.mee.2007.10.002

Chaujar R, Kaur R, Saxena M, Gupta M (2009) TCAD assessment of gate electrode workfunction engineered recessed channel (GEWE-RC) MOSFET and its multi-layered gate architecture, part II: analog and large signal performance evaluation. Superlattices Microstruct 46:645. https://doi.org/10.1016/j.spmi.2009.07.027

Colinge JP, Kranti A, Yan R, Lee CW, Ferain I, Yu R, Akhavan ND, Razavi P (2011) Junctionless nanowire transistor (JNT): properties and design guidelines. Solid State Electron 65–66:33. https://doi.org/10.1016/j.sse.2011.06.004

De Marchi M, Sacchetto D, Zhang J, Frache S, Gaillardon PE, Leblebici Y, De Micheli G (2014) Top-down fabrication of gate-all-around vertically stacked silicon nanowire fets with controllable polarity. IEEE Trans Nanotechnol 13:1029. https://doi.org/10.1109/TNANO.2014.2363386

Dey RS, Raj CR (2013) Redox-functionalized graphene oxide architecture for the development of amperometric biosensing platform

Dutta R, Kundu S (2012) Effect of polysilicon gate doping concentration variation on MOSFET characteristics. In: 2012—5th international conference on computers and devices for communication, vol 3 (IEEE, 2012), pp 3–6. https://doi.org/10.1109/CODEC.2012.6509245

Elaty DA, Fakhry MG (2019) A smart drug delivery silica nanoparticles system for diabetes mellitus disease treatment. Acta Sci Pharm Sci 3:52. https://doi.org/10.31080/asps.2019.03.0444

Gao J, Huang W, Chen Z, Yi C, Jiang L (2019) Simultaneous detection of glucose, uric acid and cholesterol using flexible microneedle electrode array-based biosensor and multi-channel portable electrochemical analyzer. Sens Actuators B Chem. https://doi.org/10.1016/j.snb.2019.02.020

Garcia Ruiz F, Tienda-Luna IM, Godoy A, Donetti L, Gamiz F (2009) Equivalent oxide thickness of trigate SOI MOSFETs with high-κ insulators. IEEE Trans Electron Devices 56:2711. https://doi.org/10.1109/TED.2009.2030713

Georgiev VP, Mirza MM, Dochioiu AI, Adamu-Lema F, Amoroso SM, Towie E, Riddet C, Maclaren DA, Asenov A, Paul DJ (2017) Experimental and simulation study of silicon nanowire transistors using heavily doped channels. IEEE Trans Nanotechnol 16:727. https://doi.org/10.1109/TNANO.2017.2665691

Getnet M, Chaujar R (2022) Sensitivity analysis of biomolecule nanocavity immobilization in a dielectric modulated triple-hybrid metal gate-all-around junctionless NWFET biosensor for detecting various diseases. J Electron Mater. https://doi.org/10.1007/s11664-022-09466-1

Getnet Yirak M, Chaujar R (2020) TCAD analysis and modelling of gate-stack gate all around junctionless silicon NWFET based bio-sensor for biomedical application. In: Proceedings of 2nd international conference on VLSI device, circuit and system, VLSI DCS 2020, pp 18–19. https://doi.org/10.1109/VLSIDCS47293.2020.9179866

Goel A, Rewari S, Verma S, Gupta RS (2019) Dielectric modulated triple metal gate all around MOSFET (TMGAA)for DNA biomolecule detection. In: 2018 IEEE electron devices Kolkata conference (EDKCON), vol 1. IEEE, pp 337–340. https://doi.org/10.1109/edkcon.2018.8770406

Gupta N, Chaujar R (2015) Impact of device parameter variation on RF performance of gate electrode workfunction engineered (GEWE)-Silicon Nanowire (SiNW) MOSFET. https://doi.org/10.1007/s10825-015-0715-z

Henry CH (2007) Electron-hole recombination. AccessScience @ McGraw-Hill, p 1

Hossain NMM, Quader S, Siddik AB, Chowdhury MIB (2017) TCAD based performance analysis of junctionless cylindrical

doublegate all around FET up to 5 nm technology node. In: 20th Int. Conf. Comput. Inf. Technol. ICCIT 2017, p 1.https://doi.org/10.1109/ICCITECHN.2017.8281858

Houl T, Low T, Xul B, Ljl M, Samudral G, Kwong DL, Box U (2004) Impact of metal gate work function on nano CMOS device performance. IEEE, p 57

Huang Z, Gao A, Chen S, Wang Y, Li T (2018) Highly sensitive juntionless nanowire transistor biosensor in detecting breast tumor marker. In: Proc. IEEE sensors 2018–Octob, p 1. https://doi.org/10.1109/ICSENS.2018.8589906

Ilatikhameneh H, Kim S, Ng K, Klimeck G (2015) Optimum high-k oxide for the best performance of ultra-scaled double-gate MOSFETs

Im H, Huang XJ, Gu B, Choi YK (2007) A dielectric-modulated field-effect transistor for biosensing. Nat Nanotechnol 2:430. https://doi.org/10.1038/nnano.2007.180

James G, Joseph S, Mathew V (2010) The influence of metal gate work function on short channel effects in atomic-layer doped DG MOSFETs. J Electron Devices 8:310

Jean YS, Wu CY (1997) The threshold-voltage model of MOSFET devices with localized interface charge. IEEE Trans Electron Devices 44:441. https://doi.org/10.1109/16.556154

Kale AS, Nemeth W, Harvey SP, Page M, Young DL, Agarwal S, Stradins P (2018) Effect of silicon oxide thickness on polysilicon based passivated contacts for high-efficiency crystalline silicon solar cells. Sol Energy Mater Sol Cells 185:270. https://doi.org/10.1016/j.solmat.2018.05.011

Kumar A, Tripathi MM, Chaujar R (2018) Comprehensive analysis of sub-20 nm black phosphorus based junctionless-recessed channel MOSFET for analog/RF applications. Superlattices Microstruct 116:171. https://doi.org/10.1016/j.spmi.2018.02.018

Kumar A, Gupta N, Tripathi MM, Chaujar R (2020) Analysis of structural parameters on sensitivity of black phosphorus junctionless recessed channel MOSFET for biosensing application. Microsyst Technol 26:2227. https://doi.org/10.1007/s00542-019-04545-6

Kumar K et al (2020) Design and optimization of junctionless-based devices with noise reduction for ultra-high frequency applications. Appl Phys A Mater Sci Process. https://doi.org/10.1007/s00339-020-04092-2

Lee J-H, Chae E-J, Park S, Choi J-W (2019) Label-free detection of γ-aminobutyric acid based on silicon nanowire biosensor. Nano Converg. https://doi.org/10.1186/s40580-019-0184-3

Li BR, Chen CW, Yang WL, Lin TY, Pan CY, Chen YT (2013) Biomolecular recognition with a sensitivity-enhanced nanowire transistor biosensor. Biosens Bioelectron 45:252. https://doi.org/10.1016/j.bios.2013.02.009

Lin R, Lu Q, Ranade P, King T, Hu C (2002) An adjustable work function technology using Mo gate for CMOS devices. IEEE Electron Device Lett 23:49

Liu J, Wen HC, Lu JP, Kwong DL (2005) Dual-work-function metal gates by full silicidation of poly-Si with Co-Ni bi-layers. IEEE Electron Device Lett 26:228. https://doi.org/10.1109/LED.2005.844696

Localized SMW (2011) A compact model for threshold voltage of interface trapped charges, vol 58, p 567

Lu N, Dai P, Gao A, Valiaho J, Kallio P, Wang Y, Li T (2014) Label-free and rapid electrical detection of HTSH with CMOS-compatible silicon nanowire transistor arrays. ACS Appl Mater Interfaces 6:20378. https://doi.org/10.1021/am505915y

Madan J, Rishu RSG (2016) Mathematical modeling insight of hetero gate dielectric—dual material gate-GAA-tunnel FET for VLSI/analog applications. Microsyst Technol. https://doi.org/10.1007/s00542-016-2872-9

Maszara WP (2005) Fully silicided metal gates for high-performance CMOS technology. J Electrochem Soc 1:341. https://doi.org/10.1149/1.1924307

Matsukawa T, Endo K, Liu Y, Shinichi O, Ishikawa Y, Yamauchi H, Tsukada J, Ishii K, Masahara M, Sakamoto K, Suzuki E (2008) A Ta/Mo interdiffusion dual metal gate technology for drivability enhancement of FinFETs. IEEE Electron Device Lett 29:618. https://doi.org/10.1109/LED.2008.922965

Nam H, Lee Y, Park J, Shin C (2016) Study of work-function variation in high-κ/metal-gate gate-all-around nanowire MOSFET. IEEE Trans Electron Devices 63:3338. https://doi.org/10.1109/TED.2016.2574328

Parihar MS, Kranti A (2015) Enhanced sensitivity of double gate junctionless transistor architecture for biosensing applications. Nanotechnology 26:145201. https://doi.org/10.1088/0957-4484/26/14/145201

Pasalic D, Marinkovic N, Feher-turkovic L (2012) Uric acid as one of the important factors in multifactorial disorders—facts and controversies. Biochem Med 22:63

Peesa RB, Panda DK (2021) Rapid detection of biomolecules in a junction less tunnel field-effect transistor (JL-TFET) biosensor. SILICON. https://doi.org/10.1007/s12633-021-00981-0

Peesa RB, Panda DK (2022) Rapid detection of biomolecules in a junction less tunnel field-effect transistor (JL-TFET) biosensor. SILICON 14:1705. https://doi.org/10.1007/s12633-021-00981-0

Piermarini S, Migliorelli D, Volpe G, Massoud R, Pierantozzi A, Cortese C, Palleschi G (2013) Uricase biosensor based on a screen-printed electrode modified with prussian blue for detection of uric acid in human blood serum. Sens Actuators B Chem 179:170. https://doi.org/10.1016/j.snb.2012.10.090

Pott V, Moselund KE, Bouvet D, De Michielis L, Ionescu AM (2008) Fabrication and characterization of gate-all-around silicon nanowires on bulk silicon. IEEE Trans Nanotechnol 7:733. https://doi.org/10.1109/TNANO.2008.2007215

Pour GB, Aval LF (2017) Highly sensitive work function hydrogen gas sensor based on PdNPs/SiO2/Si structure at room temperature. Results Phys 7:1993. https://doi.org/10.1016/j.rinp.2017.06.026

Pradhan KP, Kumar MR, Mohapatra SK, Sahu PK (2015) Analytical modeling of threshold voltage for cylindrical gate all around (CGAA) MOSFET using center potential. Ain Shams Eng J 6:1171. https://doi.org/10.1016/j.asej.2015.04.009

Pratap Y, Kumar M, Kabra S, Haldar S, Gupta RS, Gupta M (2018) Analytical modeling of gate-all-around junctionless transistor based biosensors for detection of neutral biomolecule species. J Comput Electron 17:288. https://doi.org/10.1007/s10825-017-1041-4

Pratikhya Raut DKP, Nanda U (2023) Analytical drain current model development of twin gate TFET in subthreshold and super threshold regions. Microelectronics J. https://doi.org/10.1016/j.mejo.2023.105761

Profile SEE (2018) On basic Fourier–Bessel expansions. https://doi.org/10.3842/SIGMA.2018.035

Rahman E, Shadman A, Khosru QDM (2017) Effect of biomolecule position and Fi Ll in factor on sensitivity of a dielectric modulated double gate junctionless MOSFET biosensor. Sens Bio-Sens Res 13:49. https://doi.org/10.1016/j.sbsr.2017.02.002

Rana V, Ahmad G, Ramesh AK, Das S, Singh P (2020) Diameter-dependent piezoresistive sensing performance of junctionless gate-all-around nanowire FET. IEEE Trans Electron Devices 67:2884. https://doi.org/10.1109/TED.2020.2991140

Rechem D, Latreche S, Gontrand C (2009) Channel length scaling and the impact of metal gate work function on the performance of double gate-metal oxide semiconductor field-effect transistors. Pramana J Phys 72:587. https://doi.org/10.1007/s12043-009-0052-5

Reddy NN, Panda DK (2022) Design and investigation of dielectric modulated triple metal gate-oxide-stack Z-shaped gate horizontal pocket TFET device as a label-free biosensor. J Micromech Microeng. https://doi.org/10.1088/1361-6439/ac7773

Reddy RSNN, Panda DK (2022) Analytical modelling for surface potential of dual material gate overlapped-on-drain TFET(DM-DMG-TFET) for label-free biosensing application. AEU Int J Electron Commun 151:152215. https://doi.org/10.1016/j.aeue.2022.154225

Rony MW, Bhowmik P, Myler HR, Mondol P (2017) Short channel effects suppression in a dual-gate gate-all-around si nanowire junctionless NMOSFET. In: Proc. 9th Int. Conf. Electr. Comput. Eng. ICECE 2016, p 538. https://doi.org/10.1109/ICECE.2016.7853976

Sachdeva N, Vashishath M, Bansal PK (2018) Effect of gate work-function on gate induced drain leakage of MOSFETs. IJCEM Int J Comput Eng Manag 21:11

Saib SS, Yadav S, Rahul J, Srivastava A, Raj B (2015) Impact of HfO_2 in graded channel dual insulator double gate MOSFET. J Comput Theor Nanosci 12:10. https://doi.org/10.1166/jctn.2015.3833

Sanjay BP, Vohra A (2020) Metal gate electrode, channel and gate oxide engineering to improve DC and analog/RF performance of double-gate MOSFET for high-speed applications. Appl Phys A Mater Sci Process 126:1. https://doi.org/10.1007/s00339-020-03576-5

Sengupta S, Pandit S (2014) Study of performance scaling of 22-Nm epitaxial delta-doped channel MOS transistor. Int J Electron. https://doi.org/10.1080/00207217.2014.945194

Shreya S, Khan AH, Kumar N, Amin SI, Anand S (2020) Core-shell junctionless nanotube tunnel field effect transistor: design and sensitivity analysis for biosensing application. IEEE Sens J 20:672. https://doi.org/10.1109/JSEN.2019.2944885

Siddik AB, Hossain NMM, Quader S, Chowdhury MIB (2018) Silicon on metal technology merged with cylindrical gate all around FET for enhanced performance. In: 3rd Int. Conf. Electr. Inf. Commun. Technol. EICT 2017, p 1. https://doi.org/10.1109/EICT.2017.8275181

Sjöblom G (2006) Metal gate technology for advanced CMOS devices, ACTA

Software DS (2023) Atlas User's Manual

Soylemez S, Udum YA, Kesik M, Gündoğdu Hizliateş C, Ergun Y, Toppare L (2015) Electrochemical and optical properties of a conducting polymer and its use in a novel biosensor for the detection of cholesterol. Sens Actuators B Chem 212:425. https://doi.org/10.1016/j.snb.2015.02.045

Sze (2014) Performance investigation of dual material gate stack Schottky-barrier source/drain GAA MOSFET for analog applications. Phys Semicond Devices Environ Sci Eng 10:739. https://doi.org/10.1007/978-3-319-03002-9

Tolstov GP (1962) Fourier series, 1st edn. Dover Publications Inc, New York

Tripathi SL, Patel R, Agrawal VK (2019) Low leakage pocket junction-less DGTFET with biosensing cavity region. Turk J Electr Eng Comput Sci 27:2466. https://doi.org/10.3906/elk-1807-186

Usha C, Vimala P, Samuel TSA, Pandian MK (2020) A novel 2-D analytical model for the electrical characteristics of a gate-all-around heterojunction tunnel field-effect transistor including depletion regions. J Comput Electron 19:1144. https://doi.org/10.1007/s10825-020-01503-8

Van Zeghbroeck B (2008) Principles of semiconductor devices, vol 22

Wang P, Zhuang Y, Li C, Wang H (2011) New analytical threshold voltage model for halo-doped cylindrical surrounding-gate MOSFETs. J Semicond 32:8. https://doi.org/10.1088/1674-4926/32/7/074002

Wang X, Shahidi G, Oldiges P, Khare M (2008) Device scaling of high-performance MOSFET with metal gate high-K at 32 nm technology node and beyond. In: Int. Conf. Simul. Semicond. Process. Devices, SISPAD, p 309. https://doi.org/10.1109/SISPAD.2008.4648299

Wenga G, Jacques E, Salaün AC, Rogel R, Pichon L, Geneste F (2013) Step-gate polysilicon nanowires field effect transistor compatible with CMOS technology for label-free DNA biosensor. Biosens Bioelectron 40:141. https://doi.org/10.1016/j.bios.2012.07.001

Xuan X, Yoon HS, Park JY (2018) A wearable electrochemical glucose sensor based on simple and low-cost fabrication supported micro-patterned reduced graphene oxide nanocomposite electrode on flexible substrate. Biosens Bioelectron. https://doi.org/10.1016/j.bios.2018.02.054

Yan Q, Zhi N, Yang L, Xu G, Feng Q, Zhang Q, Sun S (2020) A highly sensitive uric acid electrochemical biosensor based on a nano-cube cuprous oxide/ferrocene/uricase modified glassy carbon electrode. Sci Rep. https://doi.org/10.1038/s41598-020-67394-8

Yoon J, Shin M, Lee T, Choi JW (2020) Highly sensitive biosensors based on biomolecules and functional nanomaterials depending on the types of nanomaterials: a perspective review. Materials (basel). https://doi.org/10.3390/ma13020299

Yuan J, Woo JCS (2005) A novel split-gate MOSFET design realized by a fully silicided gate process for the improvement of transconductance and output resistance. IEEE Electron Device Lett 26:829. https://doi.org/10.1109/LED.2005.857718

Zhang D, Liu Q (2015) Biosensors and bioelectronics on smartphone for portable biochemical detection. Biosens Bioelectron. https://doi.org/10.1016/j.bios.2015.08.037

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Uncorrected Proof

Water Practice & Technology

© 2024 The Authors

Water Practice & Technology Vol 00 No 0, 1 doi: 10.2166/wpt.2024.153

Numerical simulation of flow characteristics in nonprismatic compound channels

Vijay Kaushik [a,*], Munendra Kumar[a] and Bandita Naik[b]

[a] Department of Civil Engineering, Delhi Technological University, Delhi 110042, India
[b] Department of Civil Engineering, Methodist College of Engineering, Hyderabad 500001, India
*Corresponding author. E-mail: vijaykaushik_2k20phdce01@dtu.ac.in

VK, 0000-0002-7270-4520

ABSTRACT

The assessment of the flow characteristics of river systems is a very intricate undertaking in the development of hydraulic models for the purposes of flood control and floodplain management. Therefore, it is essential to use simulation models in order to calibrate and verify the experimental results. In this study, the Hydrologic Engineering Centre's – River Analysis System (HEC-RAS) is used to calibrate and validate the distribution of velocity and shear stress for different converging compound channels. Two separate flow regimes were assessed for validation based on experimental data obtained from converging compound channels with angles of $\theta = 5°$, 9°, and 12.38°. The projected values for two relative depths ($\beta = 0.15$ and 0.20) exhibit a similar pattern of variation as the empirical observations and are marginally lower than the recorded values. This suggests that the HEC-RAS model accurately estimates the velocity and shear stress values. The disparity between the simulated and experimental outcomes shows a discrepancy of less than 10%. Hence, the implications of our results suggest that while dealing with nonprismatic rivers, it is advisable to take into account lower values. The used methodology and the outcomes focused on problem-solving might potentially inform the development of flood control infrastructure for nonprismatic watercourses.

Key words: converging floodplains, HEC-RAS modeling, nonprismatic compound channel, shear stress distribution, velocity distribution

HIGHLIGHTS

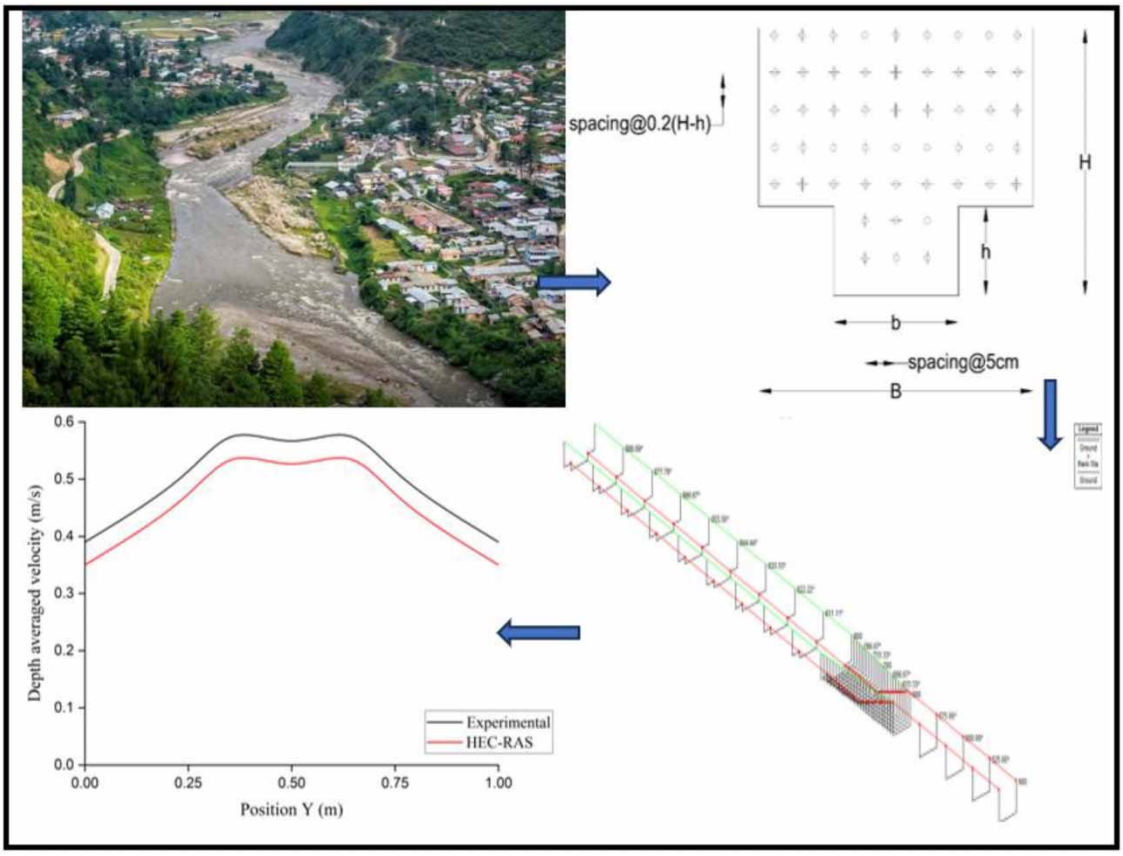
- This article presents a study that aimed to examine the flow characteristics in a nonprismatic compound channel with converging floodplains.
- The HEC-RAS software was used for this investigation.
- The findings demonstrate that the HEC-RAS models are precise in predicting the flow characteristics in nonprismatic channels, as well as replicate the experimental results.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (http://creativecommons.org/licenses/by-nc-nd/4.0/)

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

GRAPHICAL ABSTRACT



NOTATIONS

The following symbols are used in this paper:

B total width of the compound channel
b total width of the main channel
H bankfull depth
h total height of the main channel
β relative flow depth $[(H-h)/H]$
θ converging angle

ACRONYMS

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
ADV	acoustic Doppler velocimeter
ANN	artificial neural network
DCM	divided channel method
DEM	digital elevation model
GEP	gene expression programming
HEC-HMS	Hydrologic Engineering Centre's – Hydrologic Modeling System
HEC-RAS	Hydrologic Engineering Centre's – River Analysis System
LDM	lateral distribution technique
SCM	single channel method
SVM	support vector machine

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

1. INTRODUCTION

Overbank flow is the phenomenon in which water inundates the neighboring area located above the floodplain, leading to the creation of a compound channel characterized by two distinct phases. The determination of the maximum flood level is contingent upon both the flow rate and the hindrance to flow caused by bed friction and other loss mechanisms (Ervine *et al.* 2000). The disparity in velocity between floodplain and main channel flows gives rise to significant turbulence forms in the vicinity of their crossing. The formation of one or two longitudinal vertices occurs due to the presence of turbulence anisotropy at the interface and boundary, leading to the generation of secondary flow in lateral directions. According to Fernandes *et al.* (2014), many studies suggest that the presence of secondary flow in a compound channel may differ across different compartments, which is influenced by factors such as differences in bottom roughness and lateral flow depth. Consequently, there has been a recent surge of interest among academics and engineers on the need of accurately predicting flow parameters in flood situations, with the aim of mitigating damage and safeguarding human lives and property. Throughout history, several approaches and processes have been used to facilitate accurate measurement and prediction of river discharge, velocity distribution, boundary shear stress distribution, and water surface level during overbank flows. Throughout the progression of a river's flow, alterations occur in the floodplain's geometry, leading to the formation of a compound channel that may exhibit either converging or diverging characteristics. Replicating flow in a nonprismatic compound channel poses more challenges due to the increased transfer of momentum from the main channel to the floodplains. Previous studies by Sellin (1964), Myers & Elsawy (1975), Knight *et al.* (2010), and Khatua *et al.* (2012) have extensively investigated the flow characteristics of straight and meandering prismatic two-stage channels. However, there is a dearth of knowledge on nonprismatic compound channels. According to James & Brown (1977), the presence of a converging channel form leads to an increase in flow on floodplains, but the flow on expanding floodplains is diminished. The phenomenon of compound channels characterized by symmetrically narrowing floodplains has been investigated by Bousmar & Zech (2002), Bousmar *et al.* (2004), Rezaei (2006), and Rezaei & Knight (2009). These studies have identified the occurrence of additional head loss and the transfer of momentum from the main channel to the floodplains. Proust *et al.* (2006) conducted an investigation on the asymmetric geometry, specifically focusing on its quicker convergence rate. An augmented convergence angle of 22° leads to heightened levels of mass transfer and head loss. The flow behavior of skewed, two-stage converging and diverging channels was investigated by Chlebek *et al.* (2010). Rezaei & Knight (2011), Yonesi *et al.* (2013), and Naik *et al.* (2017) conducted a recent study on converging compound channels, which produced notably more accurate outcomes compared with earlier research. In their recent study, Naik *et al.* (2022) introduced a unique equation based on the gene expression programming (GEP) approach, using nondimensional variables. This equation aims to accurately forecast the water surface profile in converging compound channels. In their study, Kaushik & Kumar (2023a) used machine learning techniques to forecast the water surface profile of a compound channel with converging floodplains. Their approach included a combination of geometric and flow-related parameters. Furthermore, Kaushik & Kumar (2023b) used GEP as a research approach to construct a novel equation for compound channels with converging floodplains. The aforementioned equation is used to quantitatively assess the magnitude of the boundary shear force that is transmitted by floodplains. Kaushik & Kumar (2023c) used the support vector machine (SVM) technique to assess the water surface profile of compound channels characterized by diminishing floodplains. The use of nondimensional geometric properties facilitated the attainment of this outcome. The findings of this study indicate that the water surface profile generated by the SVM exhibits a substantial degree of concurrence with both the empirical data and the outcomes derived from previous research endeavors. In their study, Kaushik *et al.* (2023) used the Hydrologic Engineering Centre's – River Analysis System (HEC-RAS) software to conduct the calibration and validation of water surface profiles in a range of converging compound channels. The results suggest that the HEC-RAS models suggested for nonprismatic rivers are dependable in forecasting water surface profiles, demonstrating a strong ability to generalize without showing signs of overtraining. Kaushik & Kumar (2024) utilized different machine learning models to predict the shear force carried by floodplains in nonprismatic compound channel with smooth converging floodplains. The investigation conducted in this part has mostly focused on the phenomenon of flow rates. Limited research has been conducted on the flow characteristics of nonprismatic compound channels. Consequently, it is essential to engage in the modeling of flow characteristics in nonprismatic channels, as this will contribute to the improvement of flood mitigation and risk management investigations.

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

In recent decades, considerable efforts have been dedicated to the use of 2D and 3D modeling techniques in order to improve the accuracy of water level and velocity estimates in real-world river systems. Nevertheless, there are still a number of important inquiries that have not yet been addressed. Software applications such as HEC-RAS and MIKE 11 integrate calculation methodologies such as single channel method (SCM) and divided channel method (DCM). Throughout the whole segment, the SCM maintains a consistent pace. The DCM effectively partitions the cross-sectional area into several regions characterized by different flow properties, including the primary channel and adjacent floodplains. Wormleaton *et al.* (1982) argued that the SCM tends to underestimate the conveyance capacity, whereas the DCM tends to overestimate the capacity of compound channels. In subsequent years, a number of researchers have developed improved methodologies for estimating compound channel flow. Wormleaton & Merrett (1990) proposed a straightforward modification to improve the estimate of DCM, but Ackers (1992) conducted experimental adjustments to rectify the DCM. The lateral distribution technique (LDM) introduced by Wark *et al.* (1990) and the approach suggested by Shiono & Knight (1991) were developed as alternative and more advanced methodologies. Both of these strategies, resembling a quasi-2D model, rely on identical equations to ascertain the lateral velocity distribution inside the cross-sectional area. The HEC-RAS hydraulic model, created by the U.S. Army Corps of Engineers, is extensively used in both natural and man-made channels to compute the water surface elevation and other flow parameters. This model operates in both 1D and 2D dimensions, using gradually varying dimensions to account for steady and turbulent flow conditions (Brunner 2016). In addition to its primary functions, HEC-RAS also facilitates the computation of sediment transport and mobile bed dynamics, as well as the modeling of water temperature (Arcement & Schneider 1989; Brunner 2016). The use of computer programs in recent years has significantly improved the understanding and analysis of river hydraulics and associated phenomena. Leandro *et al.* (2009) provided comprehensive insights into the hydraulic models often used in open channel modeling, along with a detailed analysis of their respective strengths and limitations. Computer hydraulic models are being widely used on a global scale to facilitate flood planning in areas that are susceptible to flooding. The primary objective of using these models is to enhance comprehension of the patterns pertaining to the magnitude and frequency of floods, while simultaneously aiding in the formulation of strategies to effectively respond to potential flood events in the future (Liu & Merwade 2018). The HEC-RAS model has been used in several studies to assess flow parameters inside the primary channel and floodplain, as well as under diverse climatic conditions. In their study, Ramesh *et al.* (2000) used an optimization approach to quantify the roughness of open channel flow. The researchers included boundary conditions as constraints in their analysis. The calibration of the HEC-RAS model for river study included the use of Manning's n roughness coefficient, as documented by Hicks & Peacock (2005) and Kuriqi & Ardiçlioğlu (2018). In their study, Timbadiya *et al.* (2011) constructed an integrated hydrodynamic model using MIKE11 software. The purpose of this model was to calibrate Manning's n roughness coefficient in order to evaluate the sensitivity of flow resistance in the Tapi River located in India. In their study, Mowinckel (2011) used the HEC-RAS model to enhance the flood conveyance capacity of an artificially constructed San Jose Creek located in Goleta, California. The evaluation enabled the formulation of a new channel configuration that would more effectively handle a flood event with a recurrence interval of 100 years, while concurrently mitigating adverse impacts on the surrounding area. In their study, Parhi *et al.* (2012) used the HEC-RAS model to calibrate the channel roughness coefficient along the Mahanadi River in Odisha. The researchers have determined that a Manning's n value of 0.032 produces optimal outcomes for the Khairrmal to Munduli segment of the Mahanadi River. In their study, Boulomytis *et al.* (2017) identified that the use of Manning's n roughness coefficients in diverse hydraulic models leads to mistakes in flow forecasts pertaining to the Bashar River. The study of rivers is vital due to their frequent use in agricultural practices and hydropower generating. In order to effectively develop and deploy flood control structures and ensure optimal flow behavior, it is essential to accurately estimate the water surface elevation in the given conditions (Kuriqi *et al.* 2019). Ogras & Onen (2020) used the HEC-RAS program to analyze the hydraulic properties of the flood bed, water surface profiles, flood recurrence, and 1D floodplain study of the Tigris River. Bharath *et al.* (2021) used the HEC-RAS program to ascertain the hydrograph of outflow caused by a breach and the hydraulic conditions at crucial downstream sites. In addition, the breach outflow hydrographs are directed utilizing dynamic flood wave routing. The HEC-RAS model is further simulated with breach parameters obtained from five different empirical methodologies, and the results are then compared. Tamiru & Wagari (2021) proposed a combined approach using machine learning and HEC-RAS models to create flood inundation maps in the Baro River Basin, Ethiopia. The artificial neural network (ANN) and HEC-RAS models were combined to create a comprehensive

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

hydrological and hydraulic model that predicts both runoff and the degree of flooding. The findings indicate that incorporating a machine learning technique into the HEC-RAS model has enhanced the accuracy of conventional flood forecasting approaches by reducing spatiotemporal uncertainty. The research conducted by AL-Hussein *et al.* (2022) demonstrated that the settlements in close proximity to the mouth of the Khazir River saw a greater impact from the floods compared with those located farther downstream. The research region has shown a robust association between Hydrologic Engineering Centre's – Hydrologic Modeling System (HEC-HMS) and HEC-RAS in assessing flood risks and accurately predicting future floods. Vashist & Singh (2023) examined how changes in the upstream border data impact the extent of the flooded region. The simulated findings using a digital elevation model (DEM) with a resolution of 12.5 m show a strong agreement with the validation data and accurately match the reported flooded regions. This research demonstrates the efficacy of HEC-RAS in 2D modeling for improved management techniques.

The objective of this research is to validate the reliability of the experimental findings concerning the flow characteristics of a nonprismatic compound channel with converging floodplains. These characteristics include the stage and discharge relationship, as well as the distribution of velocity and shear stress. The approach proposed in this article uses the HEC-RAS to enhance numerical modeling. The dataset used in this study effort to complete the simulation effectively was gathered from the work of Naik *et al.* (2017), which was done on a variety of converging compound channels and provided the basis for this work. To ensure consistent evaluation and verification of the codes, same criteria, boundary conditions, cross-section data, and flow parameters were used. Ultimately, the simulated outcomes were examined and juxtaposed with preexisting empirical evidence in order to assess and authenticate the conclusions.

2. MATERIALS AND METHODS

2.1. Physical modeling

Figure 1 illustrates the flowchart outlining the methodology used in the present investigation. The experiments were conducted in the Hydraulics laboratory, which is situated under the Department of Civil Engineering at NIT, Rourkela, India. The trials were carried out inside a masonry flume with dimensions of 15 m in length, 0.9 m in width, and 0.5 m in depth. In this study, a compound cross-section was created using brick masonry, with a main channel width of 0.5 m and a depth of 0.10 m. Figure 2 provides a comprehensive depiction of the geometric attributes and cross-sectional profile of a two-stage channel. The construction of the converging portion of the channel included the use of Perspex sheet, which was employed to achieve various converging angles denoted as $\theta = 5°$, 9°, and 12.38°, respectively, while maintaining a consistent geometric configuration. The compound channel consists of a prismatic segment of 9 m in length, followed by a nonprismatic section spanning 2.28, 1.26, and 0.84 m for converging angles $\theta = 5°$, 9°, and 12.38°, respectively, with the remaining portion being the downstream section. The subcritical flow regime was seen under various circumstances in a two-stage channel with a longitudinal bed slope of 0.0011. The estimation of Manning's n value was conducted based on data obtained from both in-bank and overbank flows in the floodplains and main channel. The water supply is recirculated inside the system by the use of a pump, which transfers water from an underground sump to a reservoir situated in the experimental channel. The rectangular notch has been encircled by vertically adjustable gates and flow enhancers. The use of flume detachable tailgates facilitates the preservation of a constant flow throughout the designated test reach. The volumetric tank serves as a receptacle for the water that is discharged from the channel. Subsequently, it returns to the subterranean reservoir.

Figure 3 depicts the experimental configuration as seen from an above perspective. Figure 4 displays the plan view of the nonprismatic cross-sections of Naik *et al.* (2017) channels. Five sections were considered for the analysis: start of converging section (sec-1), 0.10 m away from sec-1 (sec-2), middle of converging section (sec-3), 0.10 m before the sec-5 (sec-4), and end of converging section (sec-5). In the context of compound channel design, it is possible to access and measure any point on the plan of the channel. The use of a mobile bridge might be employed for the purpose of gathering the measurements. The study places significant emphasis on the ratio of channel width and aspect ratio. The flow velocity at the grid positions, as shown in Figure 2, was determined via the use of a pitot static tube featuring a diameter of 4.77 mm. The determination of the order of maximum velocity for a certain flow channel was conducted using a flow detector possessing a minimal least count of 0.1°. The use of a circular scale and pointer arrangement on a flow direction sensor is employed to quantify the angle of the Pitot tube leg with respect to the longitudinal direction of the channel. The integration of the longitudinal velocity

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

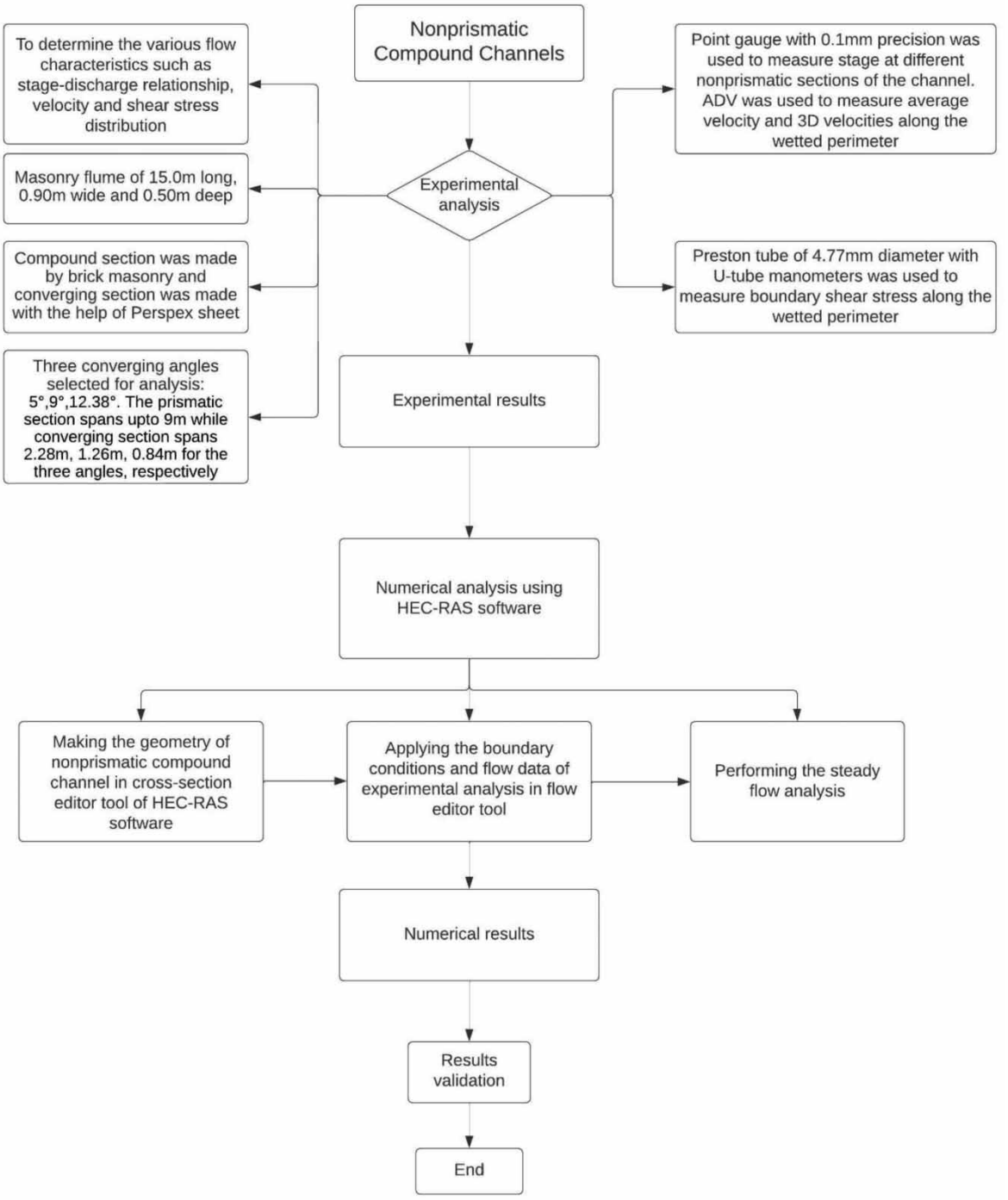


Figure 1 | Flow of methodology.

plot with the volumetric tank collection yielded a total discharge calculation that closely approximated the actual data, with a deviation of no more than $\pm 3\%$. In this work, velocity data was used in conjunction with a semi-log plot to make predictions about channel bed and wall shear forces. The calculation of boundary shear stresses was conducted by using the relationship proposed by Patel (1965) and utilizing measurements of Preston tube head differences obtained from a manometer. The correction of shear values was accomplished by conducting a comparison with the corresponding values derived using the energy gradient approach. Consequently, the findings consistently exhibited a deviation of no more than 3% from the true value. Based on the study of laboratory data, it has been shown that the accuracy of calculating tractive stresses using Pitot tubes is superior to that of using an acoustic Doppler velocimeter (ADV). One limitation to consider is the lack of reliability in detecting

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

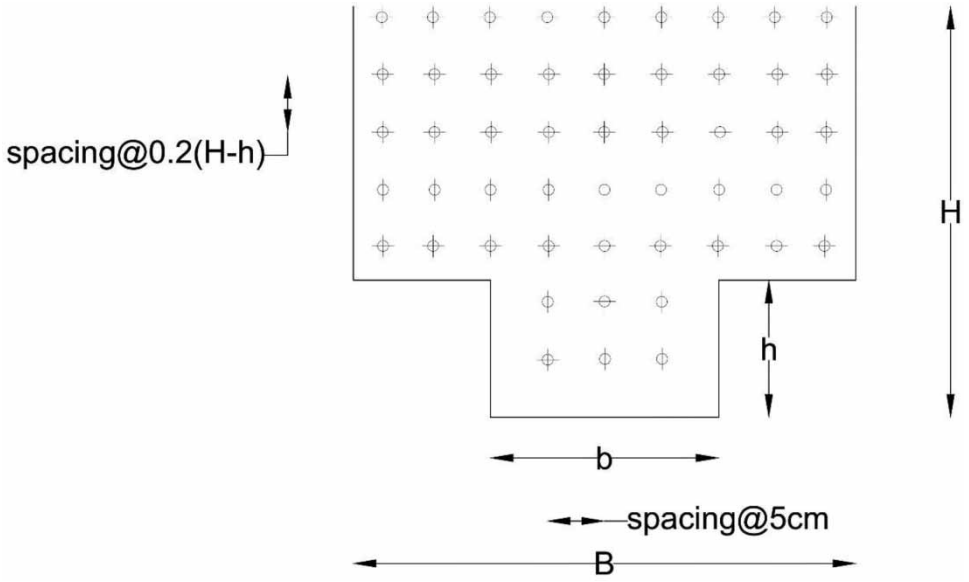


Figure 2 | Cross-section of a compound channel with geometric parameters.

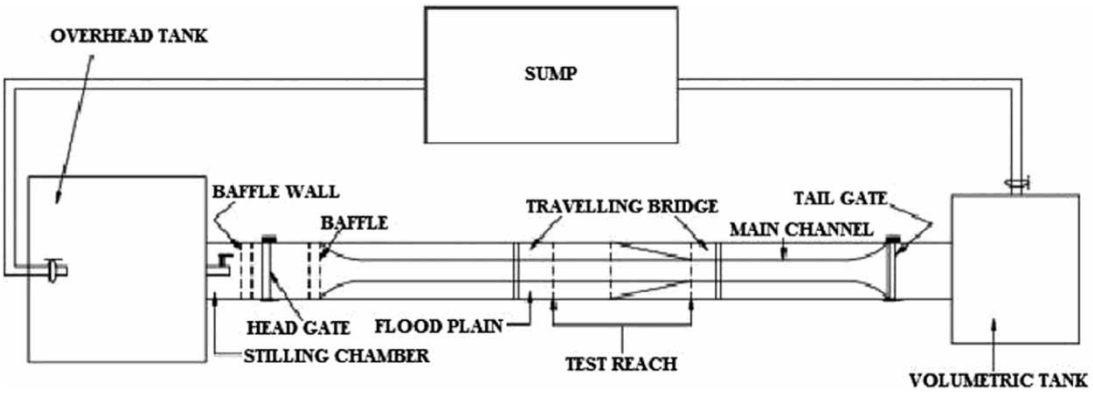


Figure 3 | Top view of the experimental setup.

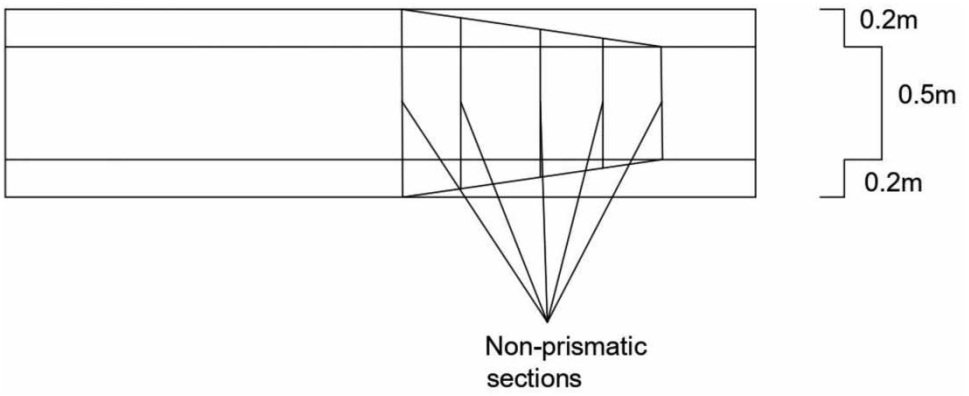


Figure 4 | Plan view of nonprismatic sections of the present study.

velocity at the border using an ADV. Moreover, ADV has considerable limitations in accurately quantifying the velocity at close proximity to the bed. The depth of penetration may reach a maximum of 5 cm below the uppermost surface. As a result, the micro-ADV down probe was unable to achieve a proximity of 5 cm from the

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

unobstructed surface. A Pitot tube was used for the purpose of quantifying the transitory reduction. The validity of this strategy was verified by the use of the energy gradient methodology, as shown by Naik *et al.* (2017).

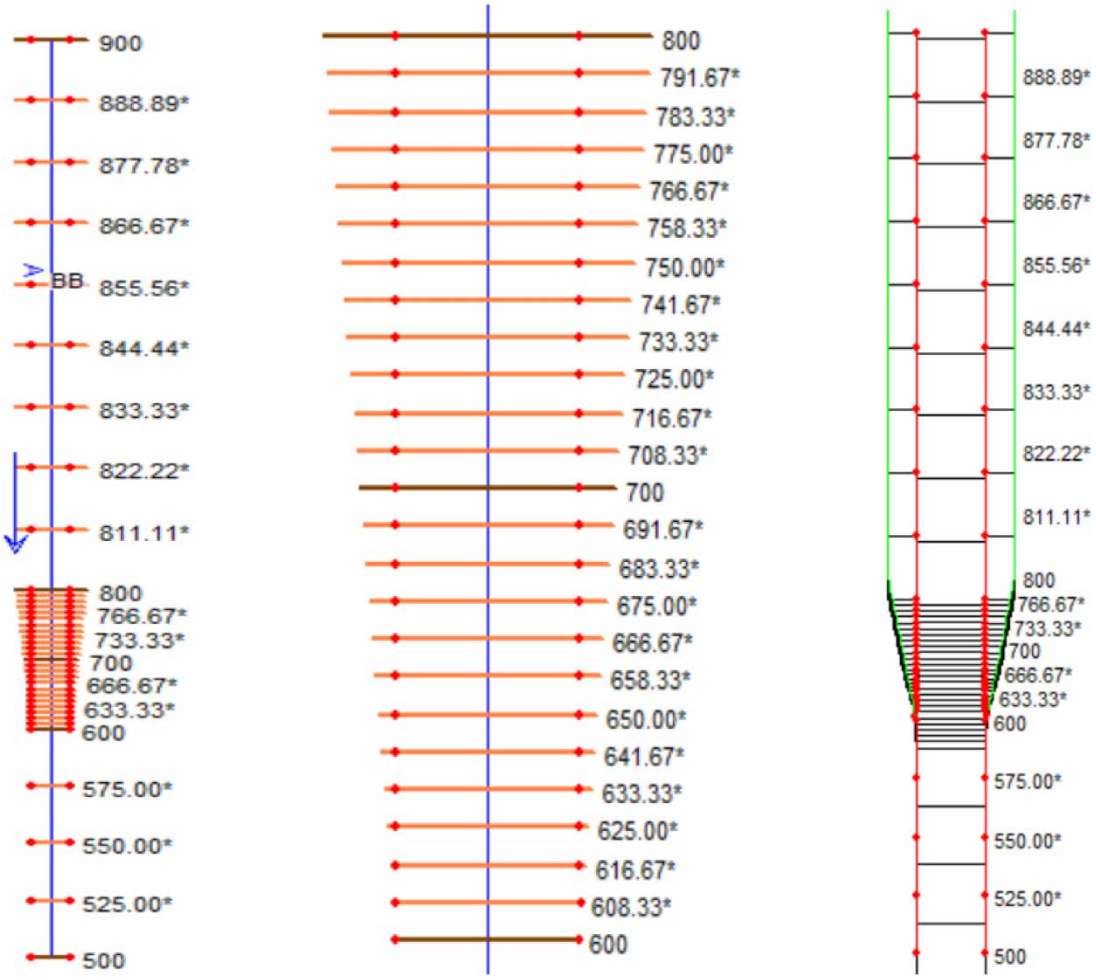


Figure 5 | Plan view of nonprismatic compound channel in HEC-RAS.

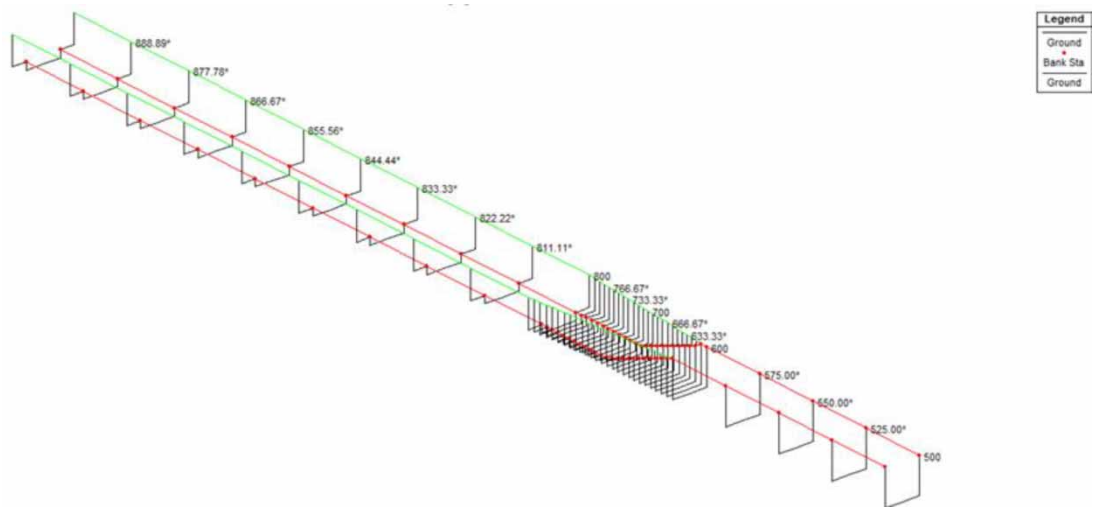


Figure 6 | 3D view of nonprismatic compound channel in HEC-RAS.

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

2.2. Numerical modeling

In order to replicate a dynamic and evolving stream, researchers choose to use the HEC-RAS model, which operates by performing calculations based on the Saint-Venant equations (1) and (2) (Brunner 2016).

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = q_l \tag{1}$$

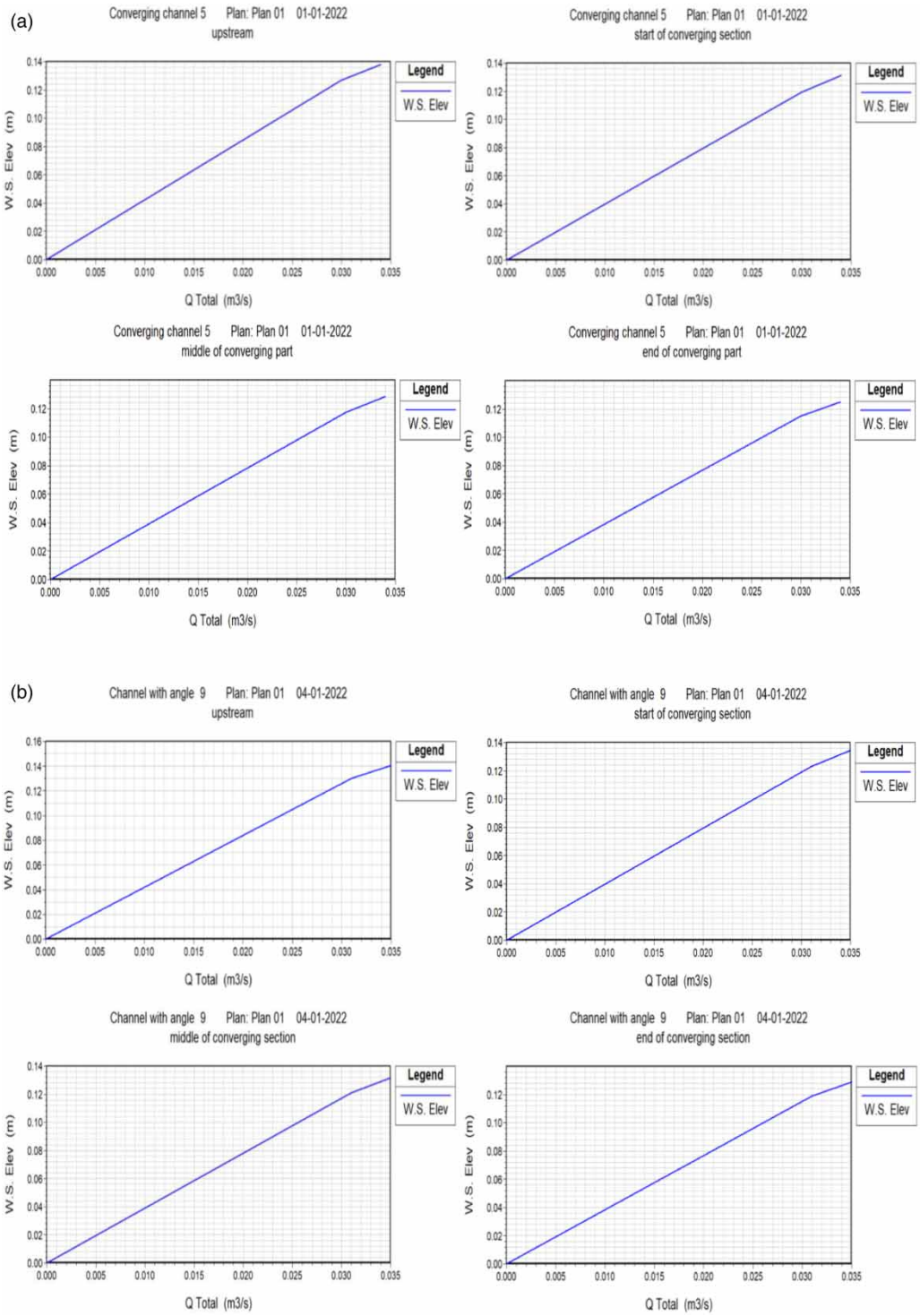


Figure 7 | Stage discharge relationship for relative depth, $\beta = 0.15$ at the upstream of the channel, as well as the start, middle, and end of the converging portion for various converging compound channels: (a) $\theta = 5°$, (b) $\theta = 9°$, and (c) $\theta = 12.38°$. (*continued.*).

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

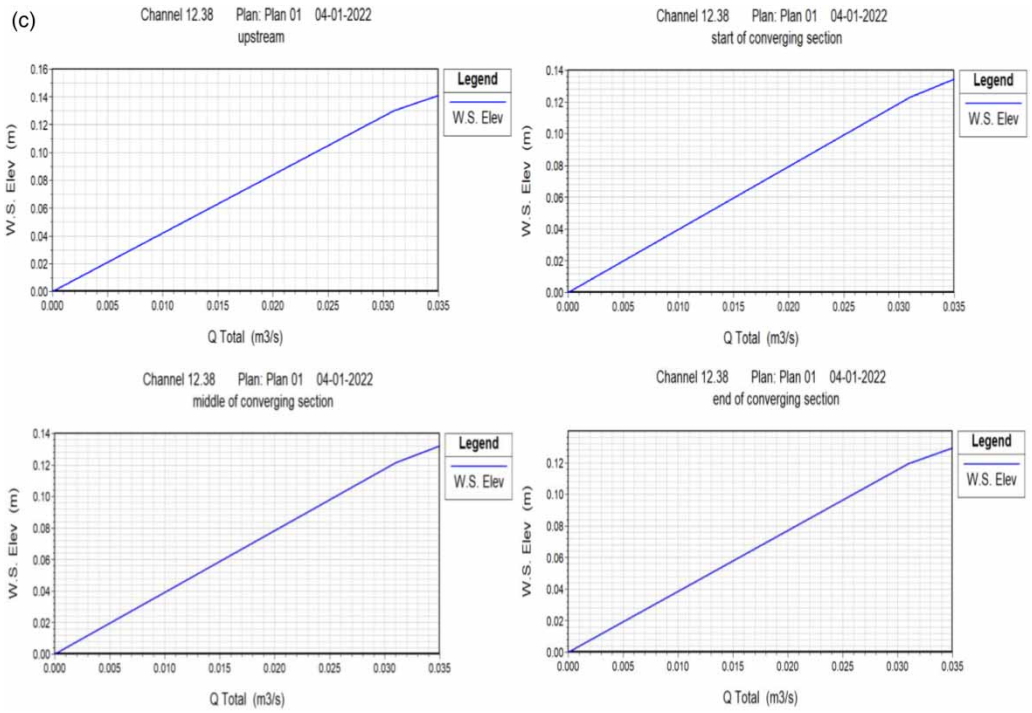


Figure 7 | Continued.

$$\frac{\partial Q}{\partial t} + \frac{\partial\left(\frac{Q^2}{A}\right)}{\partial x} + gA\frac{\partial H}{\partial x} + gA(S_o - S_f) = 0 \tag{2}$$

Equations (1) and (2) correspond to the continuity and momentum equations, respectively. In these equations, H represents the elevation of the water level, t and x represent the temporal and longitudinal coordinates, q_l represents the lateral inflow (which is zero in our case), S_o represents the bed slope of the channel, and S_f denotes the slope of the total energy line or friction slope. The equations were calculated using a finite-difference method that relied on a four-point implicit box, as described by Brunner (2016). Despite the inherent constraints of the finite-difference principle in effectively transitioning between the subcritical and supercritical flow regimes, the use of this principle requires the development of distinct solution strategies for each flow state. Nevertheless, it is possible to circumvent the aforementioned limitation by using the mixed-flow regime option of the HEC-RAS model. This enables HEC-RAS to suggest potential remedies for sub-zones within the river reach (Hicks & Peacock 2005; Timbadiya *et al.* 2011; Brunner 2016). To conduct the hydraulic simulation, many input data are necessary, including roughness, channel geometry, boundary conditions, and discharge data (Hicks & Peacock 2005; Brunner 2016). Manning's n values were included during the simulation phase, using the values used in the physical modeling. As a result of the persistent and consistent flow conditions observed at the specific section of the river under investigation, the downstream boundary condition was restored to its original state. Figures 5 and 6 illustrate the plan and 3D views of the converging compound channel geometry in the HEC-RAS. These visual representations were generated with laboratory geometry data and include the most often seen cross-sections. Figure 5 displays the upstream and downstream of the channel via the representation of Stations 900 and 500, respectively. Stations 800, 700, and 600, however, symbolize the initial, intermediate, and final segments of the converging portion, correspondingly. All stations located in the straight and converging parts of the two-stage channel, except the endpoints, may be classified as interpolated stations. The predicted velocity and shear stress obtained by HEC-RAS was afterwards compared with the observed values.

3. RESULTS AND DISCUSSION

Figure 7 illustrates the connection between stage and discharge for a certain relative depth, $\beta = 0.15$, at several locations along the converging part of distinct channels. These locations include the upstream of the channel,

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

as well as the start, middle, and end of the converging portion. The converging compound channels under consideration have different convergence angles, namely $\theta = 5°$, 9°, and 12.38°. The increase in discharge results in a corresponding rise in flow depth. However, it is important to note that beyond the bankfull depth, there is a little decrease in the rate of increment. The decrease in flow depth transpires as one proceeds from the start to the end of the converging section due to the narrowing of the channel, while maintaining the same discharge. Furthermore, it has been shown that an increase in the angle of floodplain convergence at a given stage results in a corresponding increase in the flow rate. This decrease in flow depth may be attributed to the combined influence of momentum and channel geometry. As the amount of water flowing out increases, the higher speed of the flow creates more force, which leads to the wearing away of the sides and bottom of the channel. In a converging

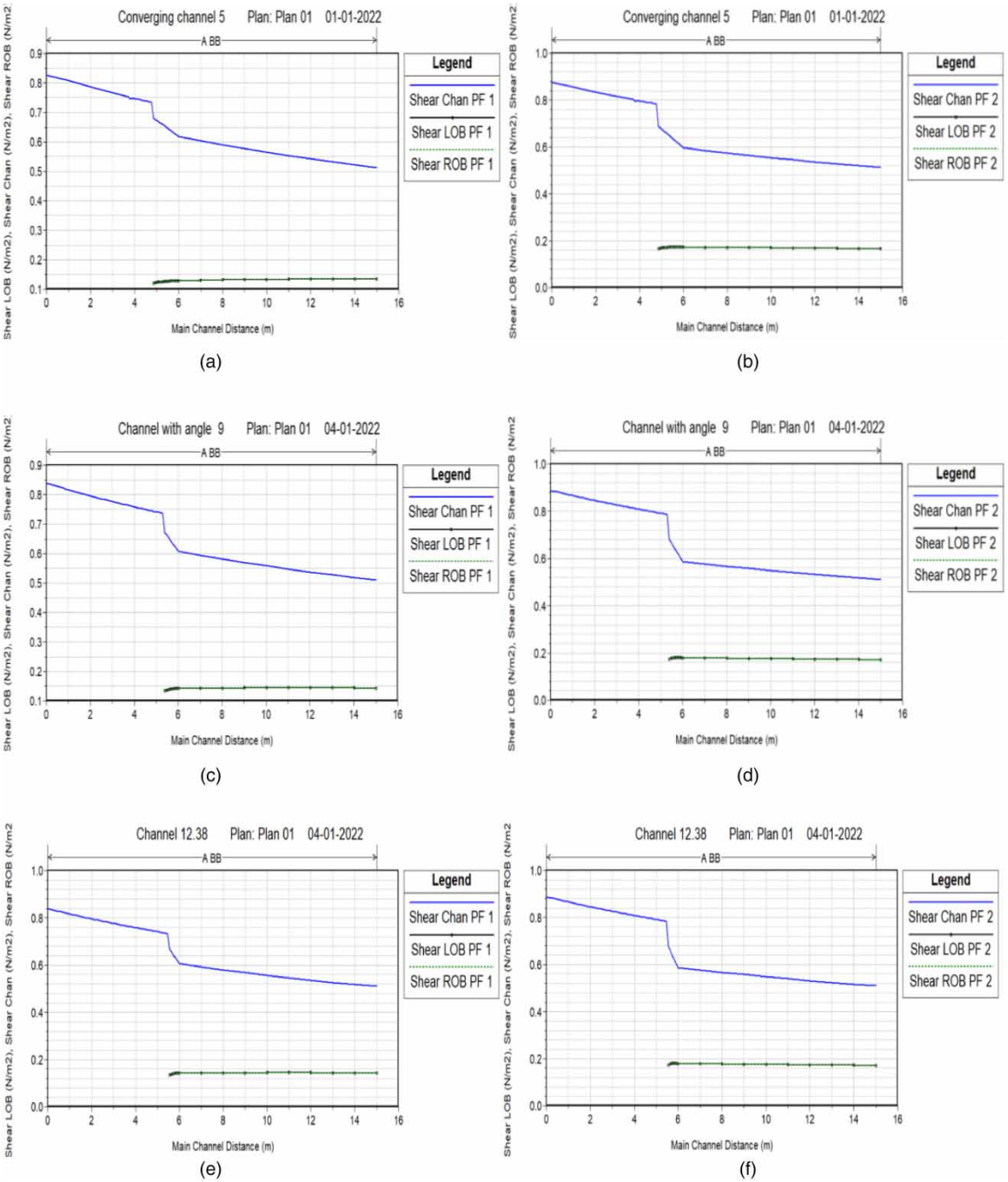


Figure 8 | Variation of average boundary shear stress with the longitudinal distance for different relative depths and converging angles: (a) $\beta = 0.15$, $\theta = 5°$; (b) $\beta = 0.20$, $\theta = 5°$; (c) $\beta = 0.15$, $\theta = 9°$; (d) $\beta = 0.20$, $\theta = 9°$; (e) $\beta = 0.15$, $\theta = 12.38°$; and (f) $\beta = 0.20$, $\theta = 12.38°$.

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

compound channel, when the main channel narrows as it meets floodplains, the expanded flow area spreads out the greater amount of water being carried, leading to a reduction in water level. The reason for this is that a channel has the capacity to hold a larger amount of water without causing a substantial increase in water level.

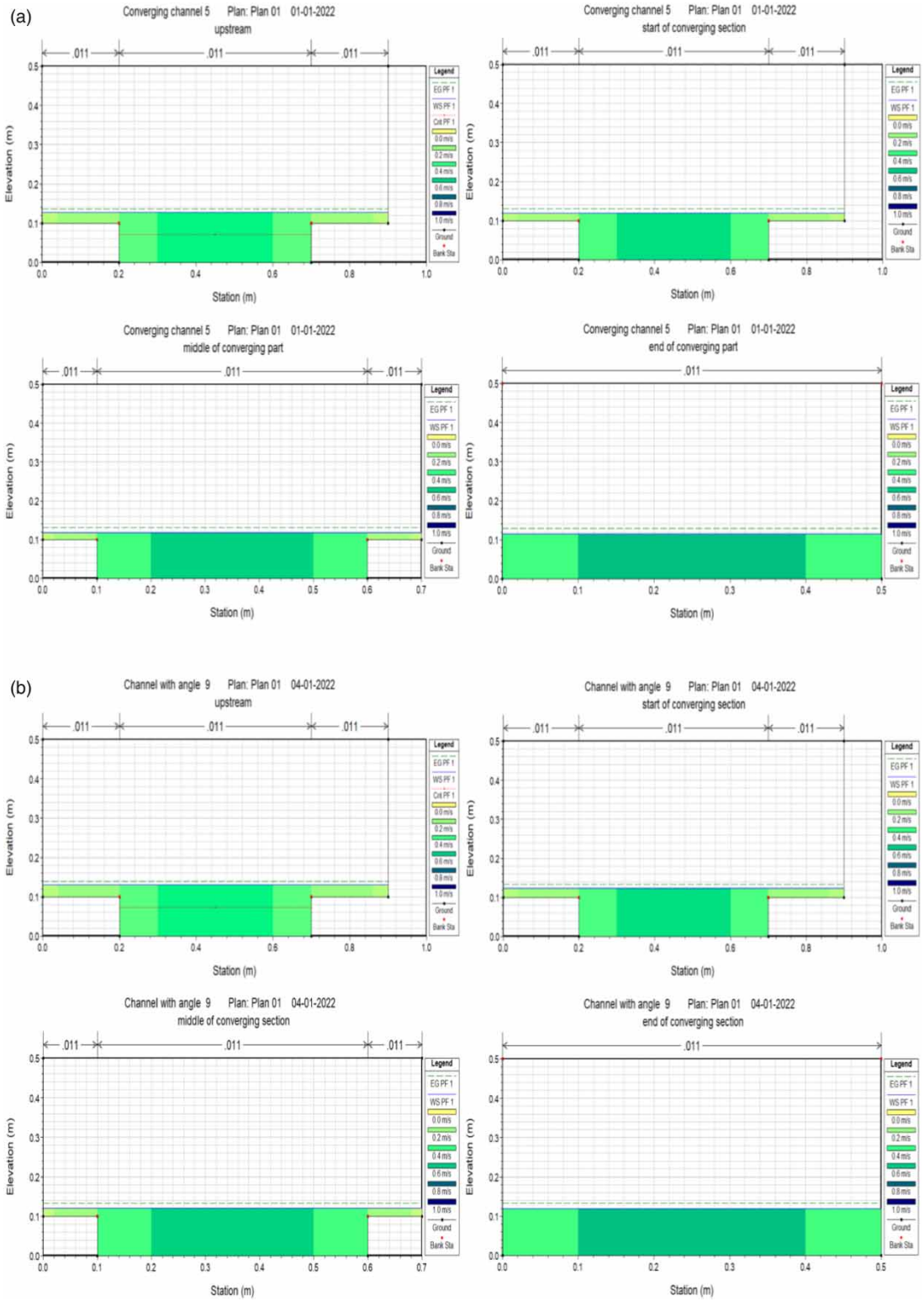


Figure 9 | Depth-averaged velocity distribution across the width of channel for relative depth, $\beta = 0.15$ at the upstream of the channel, as well as the start, middle, and end of the converging portion for various converging compound channels: (a) $\theta = 5°$, (b) $\theta = 9°$, and (c) $\theta = 12.38°$. (*continued.*).

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

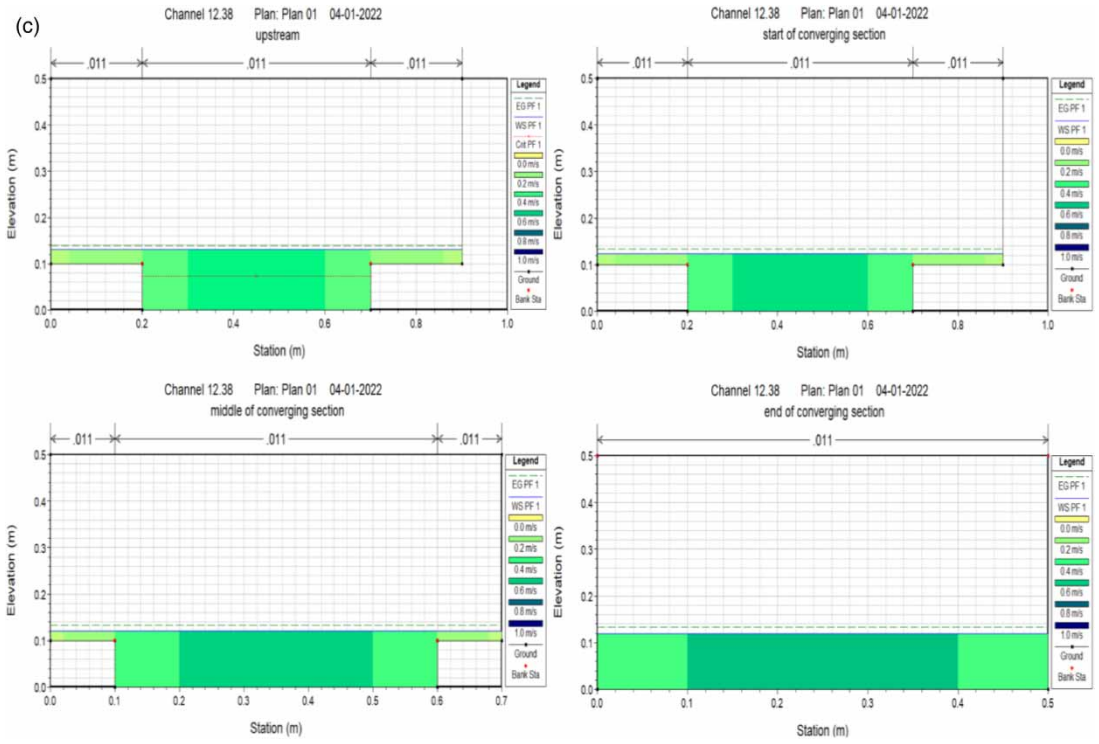


Figure 9 | Continued.

Therefore, the stage reduces as the flow spreads out across the channel cross-section, while still preserving hydraulic equilibrium. Figure 8 represents the relationship between average boundary shear stress and longitudinal distance for different relative depths ($\beta = 0.15$, 0.20) in converging compound channels with varying angles ($\theta = 5°$, 9°, 12.38°). The boundary shear stress exhibits a linear increase with longitudinal distance up to the prismatic section. However, in the converging region of the channel, a sudden spike in shear stress is seen with longitudinal distance, followed by a subsequent linear increment. The boundary shear stress in both the main channel and floodplains rises with longitudinal distance as the relative depth increases, mostly owing to the transfer of momentum from the main channel to the floodplains. As the angle of convergence grows, there is a transition in the shear stress at the beginning of the converging segment from an abrupt shift to a sharp change, followed by a linear increase as the longitudinal distance progresses. The primary contributor of this phenomenon is the converging geometry of the channel, which creates friction in the flow. This friction intensifies as the angle of convergence of the channel rises.

Figures 9 and 10 depict the velocity distribution across the width of the channel for different relative depths ($\beta = 0.15$, 0.20) at various locations, namely the upstream, start, middle, and end of the converging portion of the channel. These figures correspond to different converging compound channels with angles of convergence (θ) equal to 5°, 9°, and 12.38°. From these figures, it has been observed that the depth-averaged velocity distributions are reasonably symmetric in all sections and gradually increase from upstream to end of the converging section. The velocity in the main channel is found to be greater than that in the floodplains, and it progressively rises in the downstream direction owing to the flow acceleration from the converging geometry of the channel. At the higher relative depths, the increase in the center line velocity is found to be more as the flow progresses downstream. The velocity in the floodplains decreases and reaches to minimum at the channel boundaries at lower relative depths. While, at higher relative depths, velocity increases in the interface region of main channel and floodplains due to momentum transfer from main channel to floodplains and then, reaches to minimum at the floodplain ends. The presence of converging geometry in a channel leads to heightened turbulence as the angle of convergence becomes larger. Consequently, this increased turbulence results in a greater variation in velocity distribution across various cross-sections of the channel. Figures 11 and 12 illustrate the comparison between experimental and simulated data obtained via the use of HEC-RAS. These data pertain to the distribution of boundary shear stress and depth-averaged velocity for relative depth, $\beta = 0.15$ in various

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

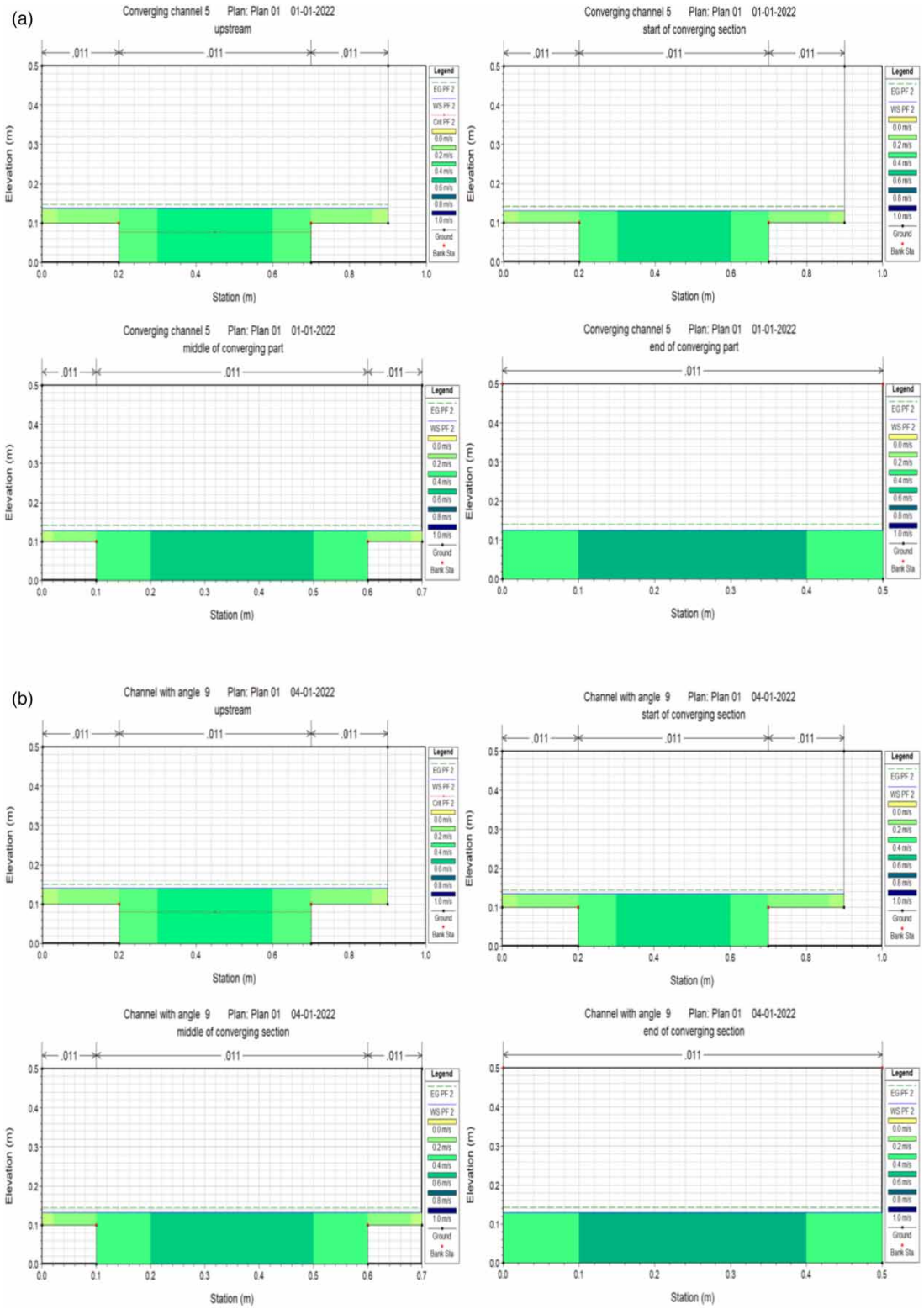


Figure 10 | Velocity distribution across the width of channel for relative depth, $\beta = 0.20$ at the upstream of the channel, as well as the start, middle, and end of the converging portion for various converging compound channels: (a) $\theta = 5°$, (b) $\theta = 9°$, and (c) $\theta = 12.38°$. (*continued*.).

compound channels with narrowing floodplains. The simulated values exhibit a consistent pattern of fluctuation that aligns with the observed trend in the experimental data. The simulated values for shear stress and velocity distribution exhibit a modest underestimation compared with the actual values. Moreover, this difference is

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

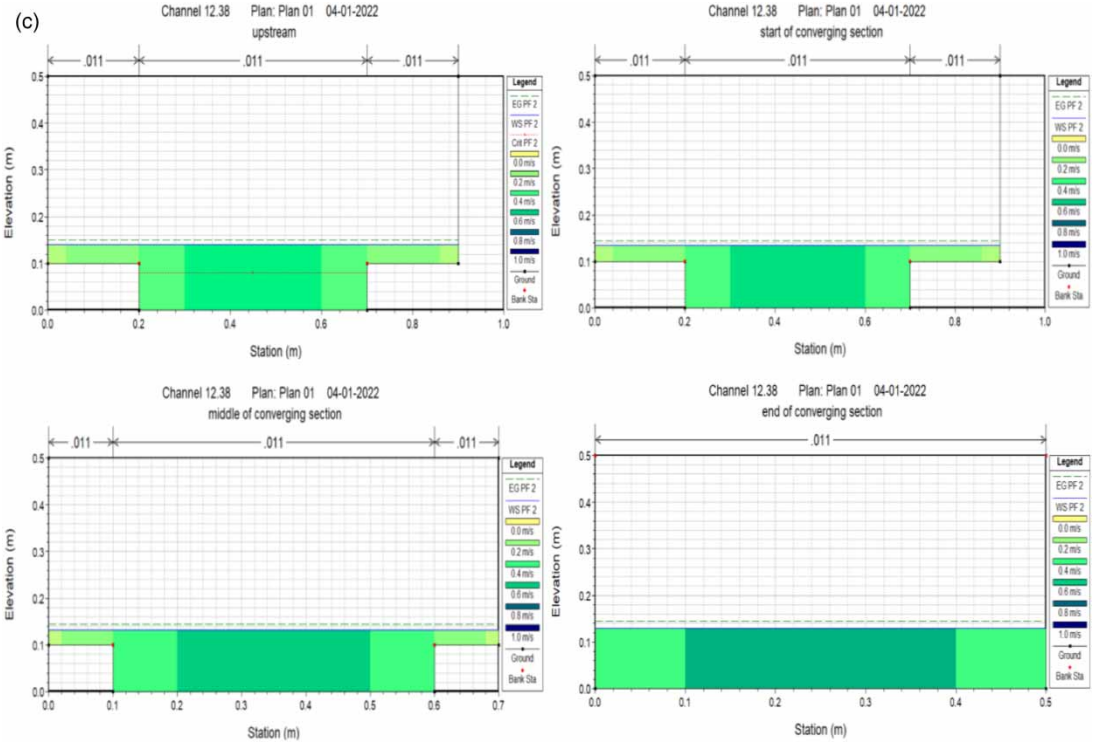


Figure 10 | Continued.

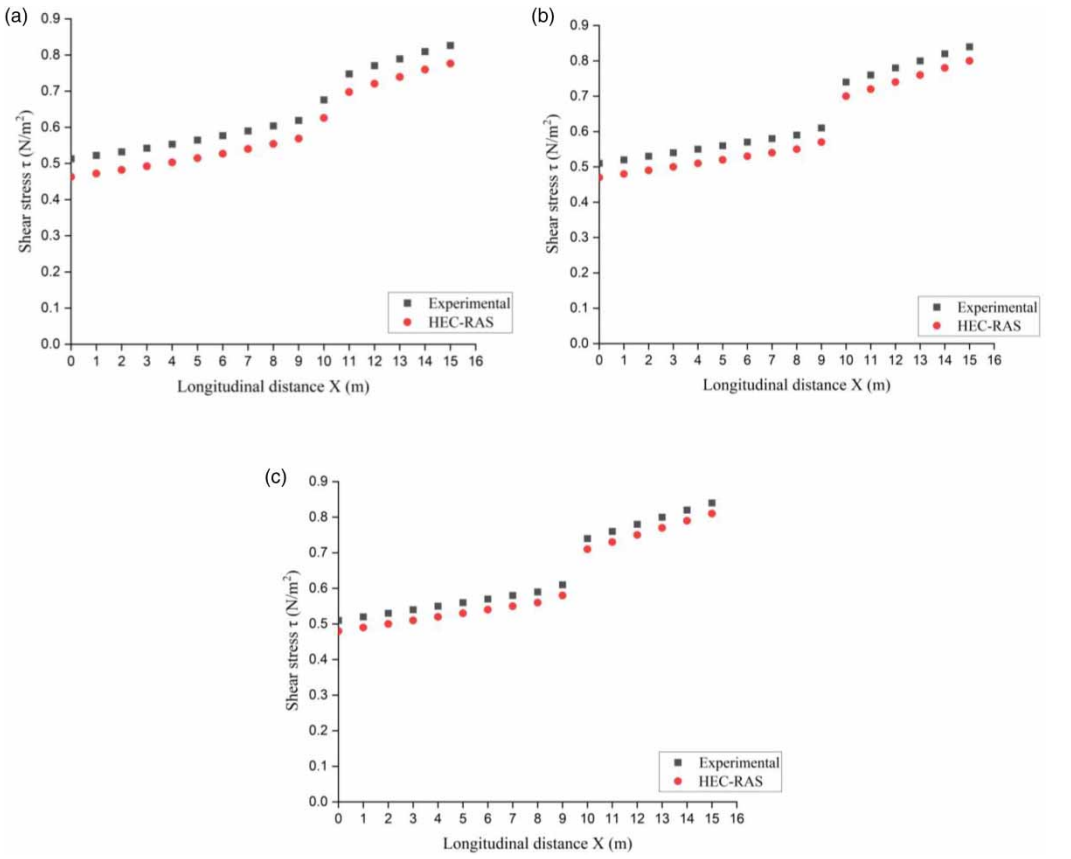


Figure 11 | Comparison of experimental and HEC-RAS simulated values of average boundary shear stress for different converging compound channels: (a) $\theta = 5°$, (b) $\theta = 9°$, and (c) $\theta = 12.38°$.

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

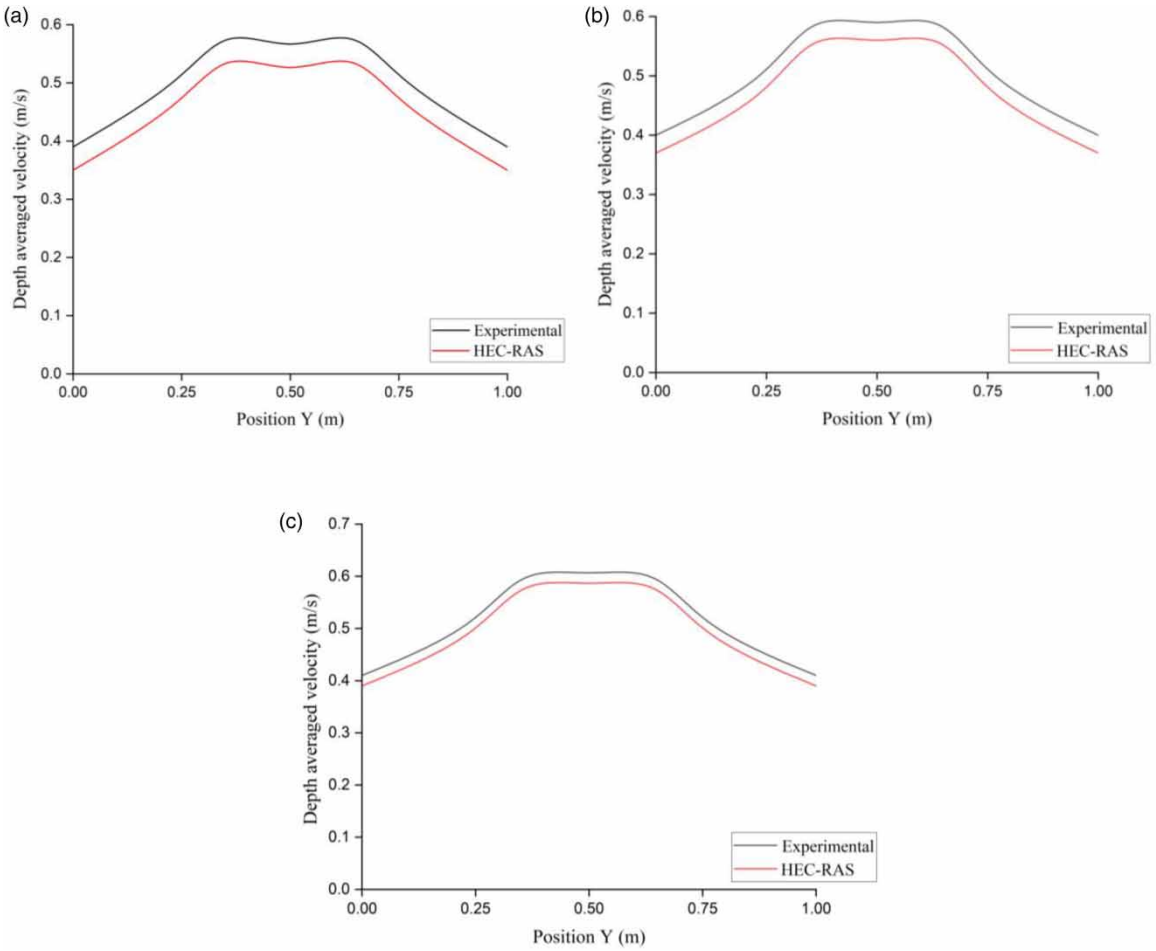


Figure 12 | Comparison of experimental and HEC-RAS simulated values of depth-averaged velocity across the width of the channel for different converging compound channels: (a) $\theta = 5°$, (b) $\theta = 9°$, and (c) $\theta = 12.38°$.

more pronounced for shallower relative depths, whereas it is less prominent for deeper relative depths. As the angle of floodplain convergence increases, the variability in estimated values decreases.

4. CONCLUSIONS

The HEC-RAS models were used in an effort to replicate the velocity and shear stress distribution in a nonprismatic compound channel with converging floodplains. In order to achieve the intended objective, an examination was conducted on two different relative depths, namely $\beta = 0.15$ and 0.20, as well as three converging channels with angles of $\theta = 5°$, 9°, and 12.38°. The increase in discharge leads to a corresponding rise in flow depth up to the bankfull depth. However, once the bankfull depth is reached, the interaction and momentum transfer between the main channel and floodplains cause a little reduction in depth at all converging angles. The longitudinal orientation of the channel segment exhibits an increase in boundary shear stress. The phenomenon of a significant rise at the beginning of the converging phase of a channel geometry is seen due to convergence. This trend is also observed for increasing relative depths and varying floodplain convergence angles. The velocity in the main channel exhibits a slight increase compared with the velocity in the floodplains. This increase is observed to be more pronounced in the downstream direction, which can be attributed to the convergence and momentum shift phenomenon. Furthermore, this pattern of variation holds true for higher relative depths and varying floodplain converging angles. The anticipated values for boundary shear stress and velocity in HEC-RAS demonstrate a little disparity in comparison to the empirical findings, while still adhering to a regular pattern. The observed discrepancy between the experimental and simulated values is determined to be less than 10%.

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

This research utilizes data from many cross-sections both upstream and downstream of the laboratory channels in order to demonstrate the capability of the HEC-RAS model in evaluating the distribution of velocity and shear stress in various converging compound channels. Consequently, irrespective of the specific model used for conducting the analysis, it is essential to possess a comprehensive comprehension of the model's requisite parameters, as well as a thorough evaluation of the simulation findings, prior to continuing further. The anticipated outcomes are influenced by many factors, including the slope of the bed, the resistance to flow, the presence of secondary currents, localized differences in channel form, and the two-dimensional effects resulting from the curvature of the channel. Comprehending the flow characteristics in nonprismatic compound channels is essential for advancing river hydraulics due to many factors. Nonprismatic channels have diverse forms and cross-sectional areas along their length, which greatly impact flow characteristics such as velocity dispersion, turbulence, and sediment transport. Engineers may enhance their ability to forecast the behavior of rivers in response to changes in flow rates, channel geometry, and sediment transport by understanding these attributes. This knowledge is valuable in designing structures such as bridges, dams, and flood control systems. Furthermore, a comprehensive understanding of nonprismatic channel flow is crucial for precisely simulating hydraulic phenomena in river systems, facilitating improved water resource management, flood prevention, and environmental preservation endeavors. HEC-RAS is a powerful tool for accurately simulating the flow behavior in irregularly shaped channels, which is essential for studying river hydraulics. It provides powerful computational skills to model intricate flow patterns, sediment movement, and hydraulic structures in nonprismatic channels. HEC-RAS incorporates variations in channel geometry, including changes in width, depth, and slope throughout the river reach, allowing for precise depiction of flow characteristics. HEC-RAS utilizes sophisticated algorithms to forecast water surface profiles, flow velocities, shear stresses, and other essential hydraulic parameters that are vital for engineering design and river management. In the context of future research, it would be intriguing to explore the economic impact that accurate evaluation of these parameters under overbank flow conditions has on the formulation of complete flood defense strategies or channel enhancement projects.

ACKNOWLEDGEMENTS

The authors would like to extend their sincere gratitude to the anonymous reviewers for their valuable contributions and diligent efforts in reviewing this manuscript.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Ackers, P. 1992 Hydraulic design of two-stage channels. *Proceedings of the ICE Water Maritime and Energy* **96**(4), 247–257.

AL-Hussein, A. A. M., Khan, S., Ncibi, K., Hamdi, N. & Hamed, Y. 2022 Flood analysis using HEC-RAS and HEC-HMS: A case study of Khazir River (Middle East – Northern Iraq). *Water* **14**(22), 3779. https://doi.org/10.3390/w14223779.

Arcement, G. J. & Schneider, V. R. 1989 *Guide for Selecting Manning's Roughness Coefficients for Natural Channels and Floodplains*, Vol. I. United States Government Printing Office, Washington, DC. https://doi.org/10.3133/wsp2339.

Bharath, A., Shivapur, A. V., Hiremath, C. G. & Maddamsetty, R. 2021 Dam break analysis using HEC-RAS and HEC-GeoRAS: A case study of Hidkal dam, Karnataka state, India. *Environmental Challenges* **5**, 100401. https://doi.org/10.1016/j.envc.2021.100401.

Boulomytis, V. T. G., Zuffo, A. C., Dalfré Filho, J. G. & Imteaz, M. A. 2017 Estimation and calibration of Manning's roughness coefficients for ungauged watersheds on coastal floodplains. *International Journal of River Basin Management* **15**(2), 199–206. https://doi.org/10.1080/15715124.2017.1298605.

Bousmar, D. & Zech, Y. 2002 Periodical turbulent structures in compound channels. In *River Flow International Conference on Fluvial Hydraulics*, Louvain-la-Neuve, Belgium, pp. 177–185.

Bousmar, D., Wilkin, N., Jacquemart, J. H. & Zech, Y. 2004 Overbank flow in symmetrically narrowing floodplains. *Journal of Hydraulic Engineering* **130**(4), 305–312.

Brunner, G. W. 2016 *HEC-RAS River Analysis System (Trans: Center HE)*. U.S. Army Corps of Engineers, Davis, CA. Available from: https://www.hec.usace.army.mil/software/hec-ras/documentat ion.aspx, 5.0 edn.

Chlebek, J., Bousmar, D., Knight, D. W. & Sterling, M. A. 2010 Comparison of overbank flow conditions in skewed and converging/diverging channels. In *River Flows International Conference*, pp. 503–511.

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

Doncker, L., Troch, P., Verhoeven, R., Bal, K., Meire, P. & Quintelier, J. 2009 Determination of the Manning roughness coefficient influenced by vegetation in the river Aa and Biebrza river. *Environmental Fluid Mechanics* **9**(5), 549–567.

Ervine, D. A., Babaeyan-Koopaei, K. & Sellin, R. H. 2000 Two-dimensional solution for straight and meandering overbank flows. *Journal of Hydraulic Engineering* **126**(9), 653–669.

Fernandes, J. N., Leal, J. B. & Cardoso, A. H. 2014 Improvement of the lateral distribution method based on the mixing layer theory. *Advances in Water Resources* **69**, 159–167.

Hicks, F. E. & Peacock, T. 2005 Suitability of HEC RAS for flood forecasting. *Canadian Water Resources Journal* **30**(2), 159–174. https://doi.org/10.4296/cwrj3002159.

James, M. & Brown, R. J. 1977 *Geometric Parameters That Influence Floodplain Flow*. U.S. Army Engineer Waterways Experimental Station, Vicksburg Miss, June, Research Report H-77.

Kaushik, V. & Kumar, M. 2023a Assessment of water surface profile in nonprismatic compound channels using machine learning techniques. *Water Supply* **23**(1), 356–378. https://doi.org/10.2166/ws.2022.430.

Kaushik, V. & Kumar, M. 2023b Sustainable gene expression programming model for shear stress prediction in nonprismatic compound channels. *Sustainable Energy Technologies and Assessments* **57**, 103229. https://doi.org/10.1016/j.seta.2023.103229.

Kaushik, V. & Kumar, M. 2023c Water surface profile prediction in non-prismatic compound channel using support vector machine (SVM). *AI in Civil Engineering* **2**, 6. https://doi.org/10.1007/s43503-023-00015-1.

Kaushik, V. & Kumar, M. 2024 Prediction of shear stress distribution in compound channel with smooth converging floodplains. *Journal of Hydrology and Hydromechanics* **72**(2), 170–184. https://doi.org/10.2478/johh-2024-0004.

Kaushik, V., Kumar, M., Naik, B. & Parsaie, A. 2023 Modeling of water surface profile in non-prismatic compound channels. *Water Practice & Technology* **18**(9), 2151–2167. https://doi.org/10.2166/wpt.2023.142.

Khatua, K. K., Patra, K. C. & Mohanty, P. K. 2012 Stage-discharge prediction for straight and smooth compound channels with wide floodplains. *Journal of Hydraulic Engineering* **138**(1), 93–99.

Knight, D. W., Tang, X., Sterling, M., Shiono, K. & McGahey, C. 2010 Solving open channel flow problems with a simple lateral distribution model. *River Flow* **1**, 41–48.

Kuriqi, A. & Ardiçlioğlu, M. 2018 Investigation of hydraulic regime at middle part of the Loire River in context of floods and low flow events. *Pollack Periodica* **13**(1), 145–156. https://doi.org/10.1556/606.2018.13.1.13.

Kuriqi, A., Pinheiro, A. N., Sordo-Ward, A. & Garrote, L. 2019 Influence of hydrologically based environmental flow methods on flow alteration and energy production in a run-of-river hydropower plant. *Journal of Cleaner Production* **232**, 1028–1042.

Leandro, J., Chen, A. S., Djordjević, S. & Savić, D. A. 2009 Comparison of 1D/1D and 1D/2D coupled (Sewer/surface) hydraulic models for urban flood simulation. *Journal of Hydraulic Engineering* **135**(6), 495–504. https://doi.org/10.1061/(ASCE)HY.1943-7900.0000037.

Liu, Z. & Merwade, V. 2018 Accounting for model structure, parameter and input forcing uncertainty in flood inundation modeling using Bayesian model averaging. *Journal of Hydrology* **565**, 138–149. https://doi.org/10.1016/j.jhydrol.2018.08.009.

Mowinckel, E. 2011 *Flood Capacity Improvement of San Jose Creek Channel Using HEC-RAS*. Calif Polytech State Univ., California.

Myers, W. R. C. & Elsawy, E. M. 1975 Boundary shears in channel with flood plain. *Journal of the Hydraulics Division, ASCE* **101**(7), 933–946.

Naik, B., Khatua, K. K., Wright, N. G. & Sleigh, A. 2017 Stage-discharge prediction for converging compound channels with narrow floodplains. *Journal of Irrigation and Drainage Engineering* **143**(8), 04017017. https://doi.org/10.1061/(asce)ir.1943-4774.0001184.

Naik, B., Kaushik, V. & Kumar, M. 2022 Water surface profile in converging compound channel using gene expression programming. *Water Supply* **22**(5), 5221–5236. https://doi.org/10.2166/ws.2022.172.

Ogras, S. & Onen, F. 2020 Flood analysis with HEC-RAS: A case study of Tigris River. *Advances in Civil Engineering* 1–13. https://doi.org/10.1155/2020/6131982.

Parhi, P. K., Sankhua, R. N. & Roy, G. P. 2012 Calibration of channel roughness for Mahanadi River, (India) using HEC-RAS model. *Journal of Water Resource and Protection* **4**(10), 847–850. https://doi.org/10.4236/jwarp.2012.410098.

Patel, V. C. 1965 Calibration of the Preston tube and limitations on its use in pressure gradients. *Journal of Fluid Mechanics* **231**, 85–208.

Proust, S., Rivière, N., Bousmar, D., Paquier, A. & Zech, Y. 2006 Flow in the compound channel with abrupt floodplain contraction. *Journal of Hydraulic Engineering* **132**(9), 958–970.

Ramesh, R., Datta, B., Bhallamudi, S. M. & Narayana, A. 2000 Optimal estimation of roughness in open-channel flows. *Journal of Hydraulic Engineering* **126**(4), 299–303.

Rezaei, B. 2006 *Overbank Flow in Compound Channels with Prismatic and Non-prismatic Floodplains*. *PhD Thesis*, University of Birmingham, UK.

Rezaei, B. & Knight, D. W. 2009 Application of the Shiono and Knight Method in the compound channel with non-prismatic floodplains. *Journal of Hydraulic Research* **47**(6), 716–726.

Rezaei, B. & Knight, D. W. 2011 Overbank flow in compound channels with non-prismatic floodplains. *Journal of Hydraulic* **137**, 815–824.

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Uncorrected Proof

Sellin, R. H. J. 1964 A laboratory investigation into the interaction between flow in the channel of a river and that of its flood plain. *LaHouille Blanche* **7**, 793–801.

Shiono, K. & Knight, D. W. 1991 Turbulent open channel flows with variable depth across the channel. *Journal of Fluid Mechanics* **222**, 617–646.

Tamiru, H. & Wagari, M. 2021 Machine-learning and HEC-RAS integrated models for flood inundation mapping in Baro River Basin, Ethiopia. *Modeling Earth Systems and Environment* **8**(2), 2291–2303. https://doi.org/10.1007/s40808-021-01175-8.

Timbadiya, P., Patel, P. L. & Porey, P. 2011 Calibration of HEC-RAS model on prediction of flood for Lower Tapi River, India. *Journal of Water Resource and Protection* **03**(11), 805–811. https://doi.org/10.4236/jwarp.2011.311090.

Vashist, K. & Singh, K. K. 2023 HEC-RAS 2D modeling for flood inundation mapping: A case study of the Krishna River Basin. *Water Practice & Technology* **18**(4), 831–844. https://doi.org/10.2166/wpt.2023.048.

Wark, J. B., Samuels, P. C. & Ervine, D. A. 1990 A practical method of estimating velocity and discharge in compound channels. In *Proc. River Flood Hydraulics*, pp. 163–172.

Wormleaton, P. R. & Merrett, D. J. 1990 An improved method of the calculation for steady uniform flow in prismatic main channel/flood plain sections. *Journal of Hydraulic Research* **28**(2), 157–174.

Wormleaton, P. R., Allen, J. & Hadjipanos, P. 1982 Discharge assessment in compound channel flow. *Journal of Hydraulics Division* **108**(9), 975–994.

Yonesi, H. A., Omid, M. H. & Ayyoubzadeh, S. A. 2013 The hydraulics of flow in nonprismatic compound channels. *Journal of Civil Engineering and Urbanism* **3**(6), 342–356. https://doi.org/10.1061/(ASCE)0733-9429(2000)126:4(299).

First received 19 September 2023; accepted in revised form 24 May 2024. Available online 8 June 2024

Downloaded from http://iwaponline.com/wpt/article-pdf/doi/10.2166/wpt.2024.153/1434356/wpt2024153.pdf
by DELHI TECHNOLOGICAL UNIVERSITY (DTU) user

Check for updates

Received: 18 November 2023 | Revised: 18 April 2024 | Accepted: 30 April 2024

IET Power Electronics

DOI: 10.1049/pel2.12723

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

Optimization of battery/ultra-capacitor hybrid energy storage system for frequency response support in low-inertia microgrid

Philemon Yegon[1,2] | **Mukhtiar Singh**[1]

[1]Department of Electrical Engineering, Delhi Technological University, Delhi, India

[2]Department of Electrical and Electronic Engineering, Kenyatta University, Nairobi, Kenya

Correspondence
Philemon Yegon, Department of Electrical and Electronic Engineering, Kenyatta University, P.O Box 43844-00100, Nairobi, Kenya.
Email: yegon.philemon@ku.ac.ke

Abstract

Modern power system networks are under statutory obligations to integrate renewable energy sources (RES). The primary reason is to meet ever-increasing energy demand and also to curtail environmental pollution by greenhouse gases. However, the higher penetration of RES has the tendency of reducing inertia of overall power system network. Consequently, frequency stability is affected and deviates beyond allowable permissible limits leading to power blackouts, load shedding, and even total system failure. To address the issues associated with reduced inertia, an optimal control of hybrid energy storage system (HESS) has been proposed. HESS is basically a combination of battery and ultracapacitor, where ultracapacitor addresses rapidly varying power component by mimicking inertia while the battery compensates long-term power variations. Thus, the HESS is effectively controlled to compensate the loss of inertia by regulating its energy flow. For the purpose of improved efficiency and better power management of the HESS, an improvised particle swarm optimization (MPSO)-based virtual inertia control design has been proposed. The proposed MPSO is utilized to tune the gains of bidirectional dc–dc converter in such a way that improves frequency nadir with faster response to transient disturbances. This proposed method is simulated in MATLAB and its merits are validated in real time using hardware in loop. On analysing of the results, it can be observed that frequency nadir is improved by 48.96% with significant reduction in rate of change of frequency in comparison to conventional particle swarm optimization.

1 | INTRODUCTION

In modern times, the energy industry has seen significant growth. Several countries around the world are facing the challenge of ever-increasing energy demand. The conventional methods of power generation are already at their peak and going to be exhausted in near future. Additionally, the traditional methods of power generation are environmentally hazardous. Fossil fuels generate greenhouse gases, which contaminate the environment and pose a threat to the living organisms inhabiting the region. In order to address these issues, renewable energy sources (RES) are being rapidly implemented in contemporary power system networks. Nevertheless, the growing integration of power electronics interfaced RES results into reduced inertia and hence, poses a significant challenge for the grid stability [1]. The reduced inertia triggers frequency instability (high rate of change of frequency (RoCoF), frequency nadir and zenith). Further, this problem is very much exaggerated in power system network confined to small geographical locations termed as microgrid.

The inertia issue in microgrid operation and control is of lot of concern and several schemes primarily based on rotational mass have been proposed. Synchronous generators operating without any load, commonly referred to as synchronous

Abbreviations: AM, analytical methods; DG, distributed generation system; FAESS and SAESS, fast-acting and slow acting energy storage system; FFR, fast frequency response; GHG, greenhouse gas; HESS, hybrid energy storage system; IGBT, insulated-gate bipolar transistor; KNDGC, Kenya National Grid Distribution Code; MG, microgrid; MPSO, improvised particle swarm optimization; NERC, North American Electric Reliability Corporation; PAM, Pinch analysis method; PID, proportional (Kp), integral (Ki), and derivative (Kd) controller; PSO, particle swarm optimization technique; PTS, power transformation system; PWM, pulse width modulation; RES, renewable energy sources; RoCoF, rate of change of frequency; RPM, Ragone plot method; SBM, search-based methods; SM, statistical methods; SOC, state of charge; VFOPID, variable fractional order PID; VSG, virtual synchronous generators; VSMs, virtual synchronous machines.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2024 The Author(s). *IET Power Electronics* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

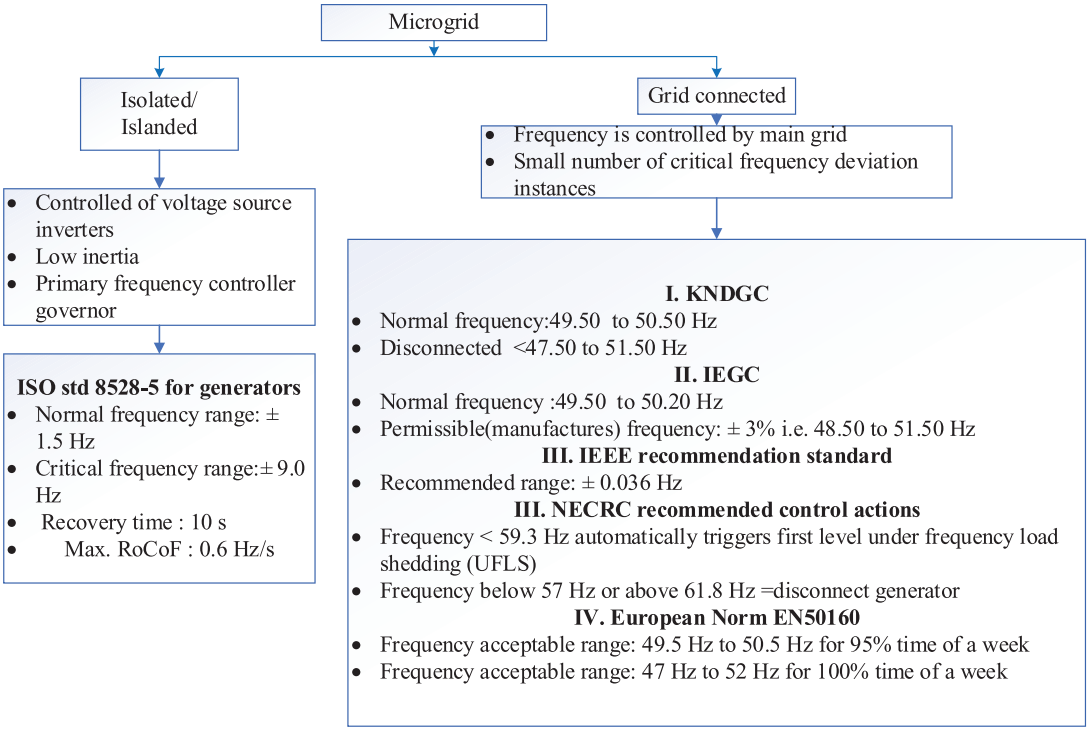


FIGURE 1 Frequency standards of microgrid [4, 5].

condensers (SC), have been widely used to compensate the loss of inertia [2]. However, the use of dedicated synchronous condenser is bulky, inefficient, and costly affair without any significant improvement in performance. To reduce the additional burden of machine, recently few schemes have been proposed where the inertia of rotating wind turbines is utilised as SC [3]. While this kind of technology seems very promising, but under variable speed operation of wind turbines, it becomes very difficult to extract the characteristics of typical SC and hence, not suitable for inertial control. The energy stored inside DC-link capacitors is also found to be very useful to overcome small transient load disturbances, but it has very limited capability heavily dependent on the size of the capacitor. Further, this kind of approach is suitable only where slight variation in DC-link voltage is allowed and not applicable to rigid DC-links comprising batteries having almost fixed voltages. Sometimes, the generating facilities in a microgrid are operated under partial-loading to keep a cushion for further increase in load. However, running a microgrid under reduced loading is extremely inefficient and increases the per unit cost of generating electricity.

Figure 1 shows frequency codes for restoration in microgrid operation prevalent in various organizations and countries.

Very recently, the energy storage systems (ESS) have been discussed widely with the intention of solving the problem of frequency instability in distributed generation system (DG) [6]. The ESS is found to be most promising for virtual synchronous machine emulation in power electronics dominant RES-based power generation. ESS having limited capacity in terms of both power and energy can be categorized on the basis of their response; rapid response ESS like flywheel, ultra-capacitors and li-ion batteries are called short-term while chemical battery (lead acid), pumped hydro storage and compressed air are known as long-term ESS. Most of the research work reported in the area of ESS power management is limited to theoretical analysis in control stability of droop, but silent about optimal sizing of ESS [7, 8]. Optimal sizing of Hybrid ESS is one of the essential ingredients not limited only to cost effectiveness but also to achieve enhanced performance out of it. Further, it is also necessary to maintain state of charge (SOC) within a given acceptable level and that is why it needs to be included in feedback control loop [9, 10]. Chu Sun proposed feedback control with additional Proportional and Integral (PI) regulator for fast-acting and slow acting energy storage system (FAESS and SAESS) to emulate synchronous generator [11]. However, in this proposal the design of PI parameters is done through the trial-and-error method. The challenge of optimal energy control was addressed by *Spyridon Chapaloglou* in [12] by controlling primary frequency through optimal methods. But, the sizing aspect of ESS is not considered in this particular article. In [13, 14], PV-battery energy storage system (BESS) is proposed and optimized using linear programming, but it did not explain effectiveness of hierarchical control nature of the systems [15, 16].

Recently, HESS-based energy management techniques have been proposed to address the inertia issue. Through the hybridization of a battery and an ultra-capacitor, fast frequency response (FFR) is achieved [17]. Inertia and damping emulation are restored using the energy recovered from them. Ultra-capacitor has high specific power density; hence, its response

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pel2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

time is rapid, that is why it is also referred to as rapid response energy storage system (RRESS). The battery has high energy density; hence, the response is slow and termed slow response energy storage system (SRESS). The idea of virtual synchronous generators (VSGs) replicated by power electronic converters is becoming increasingly popular [18]. However, problems with response time and parameter fluctuations make overall control more complex. Numerous studies have been examined to analyse the non-inertia generation effect on frequency dependability, resilience, and stability of bulk power networks [19]. In [20], the ESS has been regulated for emulating the characteristics of VSG to maintain frequency stability in low-inertia network. A typical VSG control approach incorporates the droop control loops to regulate its output active and reactive power for the better terminal voltage regulation and faster inertial response. However, sizing of HESS and determining the gains of control loops for different operating conditions is a very tedious and complex task.

An HESS-based VSG is proposed in [21, 22] where inertia constant (H) and damping constant (D) are emulated and compared with conventional parameters. Here, conventional particle swarm optimization (PSO) is used for parameter optimization which suffers from several convergence challenges. Furthermore, a low pass filter-based approach is adopted for power sharing between battery and supercapacitor without considering frequency zenith which is caused by abrupt load shedding of power supply. In [23], NN-based variable fractional order PID (VFOPID) is proposed to address reduced inertia problem in isolated microgrid which is very computationally intensive considering online updating of parameters. Since the ultracapacitor and battery have different characteristics, it is very important to analyze them independently in order to design an HESS pack. Sizing of both battery and ultra-capacitor must be optimized in such a way that it is able to handle maximum change in energy demand while keeping the voltage and frequency within permissible limits.

Although determining the size of ultracapacitor and battery forming an HESS is a typical off-line optimization problem, but optimal utilization of their capacity for a given disturbance is an online optimization issue. Therefore, the controller gains must be updated while considering the amplitude of disturbance. A typical online optimization technique has been proposed in [24] where power fluctuations are handled by updating VSG variables using the backtracking search optimization approach. Here, the damping coefficient D is updated while taking RoCoF in account whereas inertia constant H is updated depending on frequency deviation. This technique exhibits superior performance in comparison to constant variable approach but at a higher computational cost. A similar approach based on self-adapting inertia and damping constant has been proposed in [25]. In [26], a fuzzy logic-based VSG has been proposed where the parameters are updated using change in power from RES and frequency deviation only while neglecting the RoCoF. Another fuzzy logic-based inertia control approach has been proposed in [27] where both the RoCoF and the frequency deviation have been considered for inertia updation. Since fuzzy logic-based approaches are expertise dependent, the effectiveness of such an algorithm depends on accuracy of rule formulation. The methodology described in reference [28] employs a bang-bang technique based virtual inertia calculation while considering damping coefficient as constant. A PSO-based methodology has been presented in [29] where low pass filter is used to determine the power sharing of HESS. However, low pass filter has inherent disadvantage of introducing delay which affects the control performance of the system. A PSO optimized fuzzy controller is proposed in [30, 31] which is found to be marginally better than the pure fuzzy-based approach. However, writing the rules for such a heavily fluctuating system is still a big issue.

Since RES-based microgrid is a highly volatile system, the gains of its proportional (Kp), integral (Ki), and derivative (Kd) controller (PID) regulator need to be updated as per the operating conditions. Determining the inertia gain under different operating conditions is a typical optimization problem for which lots of PSO-based methods have been reported as discussed previously. But the conventional PSO is having premature convergence issue which affects the control accuracy. Therefore, in the proposed work, an improvised version of the PSO approach has been implemented for the optimum adjustment of a PID regulator gain. Here, the improved PSO is a modified version of conventional PSO by incorporating natural exponential weight updating of velocity equation to avoid it from premature convergence. Furthermore, the optimal sizing of constituent component of HESS and effective utilization of their capacity in power sharing based on their response time improves the reliability of system at significantly lesser cost. The integral squared error (ISE) of frequency deviation is taken as an objective function for the tuning of PID regulator with the aim of enhancing rate of change of the frequency, zenith and nadir while minimizing the restoration time.

Here is a summary of this paper's main contribution:

- Application of improved PSO based PID controller on bidirectional DC–DC converter interfaced HESS to enhance frequency response in microgrid under severe disturbances to ensure system stability
- Automatically tuning of PID controller parameters based on both traditional PSO and improvised PSO in the novel adaptive virtual inertia control-based HESS
- Enhancement of frequency nadir, restoration time and RoCoF using improved PSO in comparison to conventional PSO tuned PID regulator.
- Optimal techniques applied to reduce the size of ultracapacitor-based HESS as a result of reduced restoration time, frequency nadir, and energy requirement.

This article is organized in five subdivisions first section as general introduction 1, second System Configuration of Virtual Inertia regulation. The third section has Proposed Virtual Inertia Control Concept for Frequency Regulation, the fourth section has Simulation and Experimental Results and discussion with setup validation. The conclusion is in the fifth section.

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pe2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

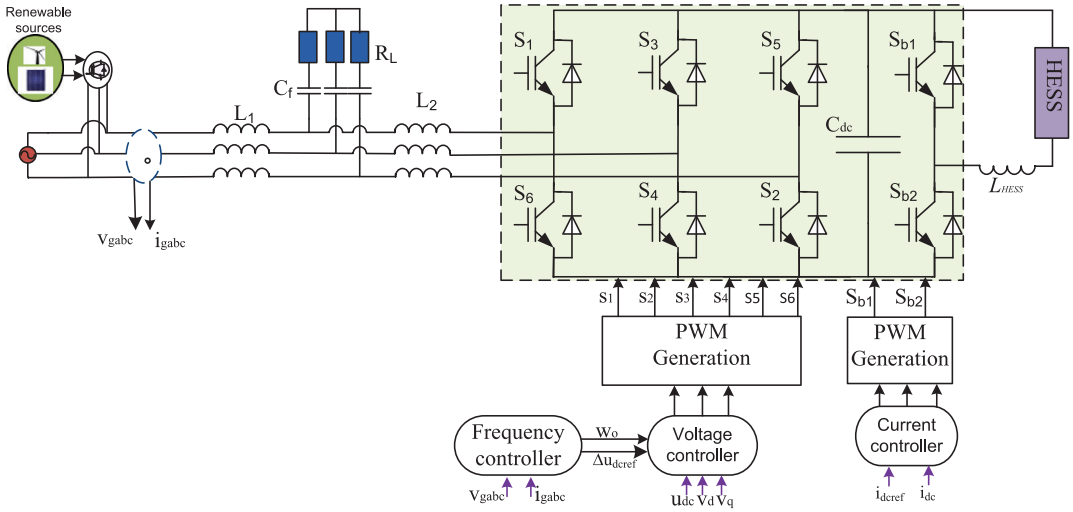


FIGURE 2 Schematic diagram of power converter in microgrid.

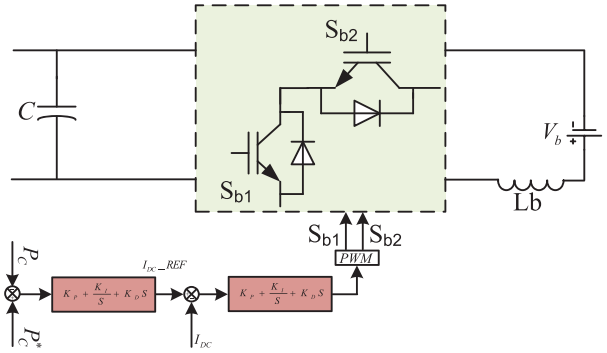


FIGURE 3 DC–DC bidirectional converter interface HESS diagram. HESS, hybrid energy storage system.

2 | SYSTEM CONFIGURATION AND CONTROL

Traditionally, synchronous generators are used to generate power and also to act as natural inertia and damping control. The turbine harnesses the stored kinetic energy to regulate the frequency. However, as renewable energy becomes more prevalent, the inherent inertia is diminishing. To accomplish this problem, the idea of VSG has been implemented [32]. In the proposed system as shown in Figure 2, a 15 MW photovoltaic (PV) generation unit (PVG), 200 mega volt amp (MVA) rated diesel generator unit (DG), wind power plant of 25 MW and battery/ultra-capacitor have been considered in the form of microgrid. Battery and ultracapacitor-based HESS has been considered to emulate the characteristics of VSG. HESS can offer active power regulation, energy management, and rapid and slow services in frequency control at a comparatively cheaper price [33]. The bidirectional DC–DC converter is used for coupling parallel combination of RRESS and SRESS to the DC-link of the grid interfacing inverter as seen in Figure 3. The common inverter uses virtual governor control similar to a traditional SG along with inertia and damping emulation as integral part of its control objectives. However, the power reference of RRESS also includes an energy recovery control. The RRESS's stored energy can be expressed as an SoC-equivalent value, such as $CV_{dc}^2/2$ for ultracapacitor.

The DC power transformation system (PTS) connected to the RRESS is responsible for maintaining DC voltage on the shared DC bus. However, the DC PTS of the SRESS is designed to adjust its power output based on the power reference from droop and integrated control loop, similarly to the turbine and governor system. Consequently, the RRESS autonomously delivers rapidly changing elements by controlling virtual inertia constants and damping constants. If the SRESS is designated as the slack bus, integral control may be used to restore the frequency. Consequently, the RRESS efficiently delivers the rapidly changing elements by effectively controlling virtual inertia and damping. The SRESS monitors the slow-fluctuating components of the VSG while detecting any changes in frequency.

In this paper, the parameters used were obtained from [34], and listed in Table 1.

2.1 | Energy variation and sizing of HESS due to frequency deviation

Sizing of ultra-capacitor is a crucial task when designing a virtual synchronous machine; it should be decided in such a way that maximum variation of energy during transient disturbance may be addressed. The greatest change in energy capacity is assessed by

$$\Delta E_{\max} = \frac{(E_{\max} - E_{nom}) + (E_{nom} - E_{\min})}{2} = E_{nom}\left(\frac{\Delta E_{\max}}{E_{nom}}\right) \quad (1)$$

By considering maximum energy changed due to frequency variation, J_{uc} inertia value which is emulated by ultra-capacitor gives; $\omega_{nom}, \omega_{zenith}$, and ω_{nadir} are nominal, zenith, and nadir frequencies respectively; and S_{rated} is the actual power of

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pel2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 System descriptions and parameter values.

Features	Symbol	Value
Battery voltage	V_{b_ref}	350 V
Grid voltage	V_{g_ref}	415 V
DC inductance	$LHESS$	5.6 mH
DC link capacitance	C_{dc}	2.82 mF
DC-link voltage	$V_{C_{dc}}$	600 V
Frequency	f_{nom}	50 Hz
Turbine time coefficient	T_{Turb}	0.5 s
Governor time coefficient	T_{Gorvn}	0.2 s
Generator inertia coefficient	H	5 s
Governor speed regulation	$Rpermin$	0.05
Damping coefficient	D	0.8

ultra-capacitor.

$$\Delta E_{\max} = \frac{J_{uc}\left(w_{nom}^2 - w_{nadir}^2\right)}{2} or \frac{J_{uc}\left(w_{zenith}^2 - w_{nom}^2\right)}{2}$$

$$J_{uc} w_{nom} \Delta w_{\max} = \frac{J_{uc} w_{nom}^2}{2} \frac{\Delta w_{\max}}{w_{nom}} = H_{uc} S_{rated} \frac{\Delta w_{\max}}{w_{nom}} \quad (2)$$

Since Equations (1) and (2) are the same, therefore equating them will give

$$E_{nom} = \frac{H_{uc} S_{rated} \frac{\Delta w_{\max}}{w_{nom}}}{\frac{\Delta E_{\max}}{E_{nom}}} \quad (3)$$

For ultra-capacitor, maximum energy change is determined by

$$\Delta E_{\max} = \frac{\left(CV_{dc,\max}^2 - CV_{dc,nom}^2\right) + \left(CV_{dc,nom}^2 - CV_{dc,\min}^2\right)}{4}$$

$$= CV_{dc,nom} \Delta V_{dc,\max}/2 = CV_{dc,nom}^2 \times \left(\frac{\Delta V_{dc,\max}}{V_{dc,nom}}\right)/2 \quad (4)$$

To determine the size of the capacitor, Equations (3) and (4) are equated with the well-designed inertia constant H_{uc};

$$CV_{dc,nom}^2 \times \left(\frac{\Delta V_{dc,\max}}{V_{dc,nom}}\right)/2 = H_{uc} S_{rated} \frac{\Delta w_{\max}}{w_{nom}}$$
$$H_{uc} = \frac{C_{dc} V_{dc}^2}{2 S_{rated}} \cdot \frac{\Delta V_{dc_\max} f_{nom}}{V_{dc} \Delta f_{\max}} \quad (5)$$

$$H_{uc} = H_{vi} \cdot \frac{\Delta V_{dc_\max} f_{nom}}{V_{dc} \Delta f_{\max}} \quad (6)$$

where $H_{vi} = \frac{C_{dc} V_{dc}^2}{2 S_{rated}}$

As is seen in Equation (6), the factors to be considered for the virtual inertia constant (H_{uc}) are: given dc-link voltage V_{dc}, maximum ratio of voltage variation $\frac{\Delta V_{dc_max}}{V_{dc}}$, minimum ratio of frequency variation $\frac{f_{nom}}{\Delta f_{\max}}$, and the rated dc-link capacitance C_{dc}.

Virtual inertia can be generated by designing a proportional controller as follows:

$$K_{vi} = \frac{\Delta V_{dc_\max}}{V_{dc}} \cdot \frac{f_{nom}}{\Delta f_{\max}} \quad (7)$$

A proportional controller for allowable deviation of both frequency and voltage is indicated by MATLAB graph as shown in Figures 4a and 4b. The coding in MATLAB is done with the following values: The frequency change $\Delta f_{\max} = 0.2, \omega_{nom} = 50$ Hz, power rated was scaled down $S_{rated} = 1000$ VA, $\frac{\Delta V_{dc,\max}}{V_{dc,nom}} = 0.15$ and $C = 2.82$ mF [35].

As is seen in Figure 4a, for inertia of 5.03823 and voltage within the acceptable threshold of $400 \pm 36V$ (in this case 418.687 V), the capacitance needed is 333.333 μF. The available retails in capacitors can be found in market store. In Figure 4b, the change in voltage is restricted to 2.52525 V which is 5.3031% and is within the acceptable limits.

2.2 | Diesel generator

The governor of the SG controls the real power. The governor's main objective is to regulate the output power of diesel engine by adjusting the mechanical power of the synchronous generator to match the difference between mechanical and electrical power.

The swing equation is as follows in Equation (8):

$$\Delta P_m(t) - \Delta P_L(t) = \frac{2Hd\Delta f(t)}{dt} + D\Delta f(t) \quad (8)$$

where ΔP_m power production varies depending on the turbine

ΔP_L change power demand
Δf frequency deviation
H inertia constant
D damping constant

Equation (8) is represented in Figure 5.

Figure 6 shows the profile of how system frequency is controlled. In any network grid frequency is among the most crucial performance indicators. Frequency degradation of the system is due to abrupt disturbance caused by generation tripping or load shedding. Frequency is regulated to prevent large (RoCoF) and high-frequency deviation beyond allowable maximum.

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pe2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

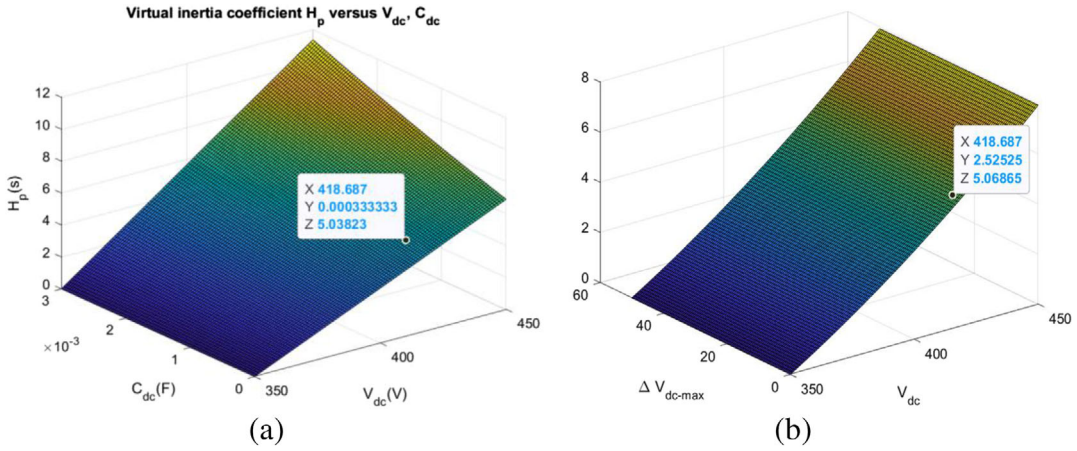


FIGURE 4 Virtual inertia coefficient (a) H_p versus $V_{dc,nom}$, C, (b) H_p versus $V_{dc,nom}$, $\Delta V_{dc,\max}$.

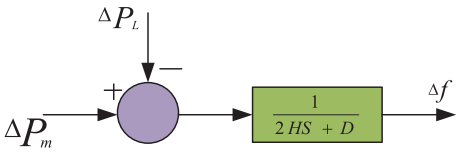


FIGURE 5 Frequency deviation model.

3 | PROPOSED VIRTUAL INERTIA CONTROL CONCEPT FOR FREQUENCY REGULATION

3.1 | Particle swarm optimization (PSO) implementation for PID controller parameters tuning

Control of HESS is essential to enhancing system stability. In this part, we will discuss how to regulate virtual inertia by determining ΔV_{dc} and Δf. This is done by utilizing a PID controller to acquire actual power reference. Here, an optimization technique is applied to automatic-tune the PID regulator. The improved PSO is a modified version of conventional PSO by incorporating natural exponential weight updating of velocity equation to avoid it from premature convergence. PSO was first introduced in 1995 by James Kennedy and Russel C. Eberhart [36]. It explains the social nature shown by animals, such as a group of birds and a group of fish. This technique is powerful, easy to implement, low computational burden, and efficient tool. It is suitable online optimization. The following formula determines the position and speed of each particle: local position p_i and best global position g_i of the particles are as follows:

$$d_i^{t+1} = d_i^t + u_i^{t+1} \tag{9}$$

$$u_i^{t+1} = wu_i^t + c_1 r_1 \left(p_i^t - d_i^t\right) + c_2 r_2 \left(g_i^t - d_i^t\right) \tag{10}$$

where r_1 and r_2 are random constant ranges (0 to 1)

c_1 and c_2 social coefficient and cognitive coefficient,
ω inertia weighting function
ωu_i^t inertia
$c_1 r_1 (p_i^t - d_i^t)$ cognitive unit
$c_2 r_2 (g_i^t - d_i^t)$ social unit

ω inertia weight, it is essential during exploitation and exploration in the search space. Well-optimized value reduces convergence period. In order to modify the inertia weight and update it dynamically, a highly successful method based on linearly decreasing inertia weight PSO was provided in reference [37] as given Equation (11).

$$w = (w_{start} - w_{end}) \times \left(\frac{T_{\max} - t}{T_{\max}}\right) + w_{end} \tag{11}$$

where ω_{start} is starting inertia weight

ω_{end} end value of inertia
$T_{\max}$ max iterations

During the searching procedure, the linear dynamic inertia weight-Particle Swarm Optimization (LDIW-PSO) encountered difficulty settling into a local optimum. Furthermore, excessively small or high values of the inertia constant may lead to an increase in the number of iterations and the time it takes for convergence, so negatively affecting the performance of the optimisation engine. In order to address this problem, we multiply the natural exponential with w_{end} to improve the equation for updating velocity, which was already introduced in our earlier publication paper [38]. Similarly, [39] have incorporated an exponential square, resulting in an additional rise in computation burden. To reduce computational burden while maintaining speed, a weight update matrix now includes an exponential component with an optimal value of $T_{\max}$ [40]. Consequently, the total time it takes for convergence lowers as the number of iterations is reduced. The adjusted Equation (12) has been modified

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pel2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

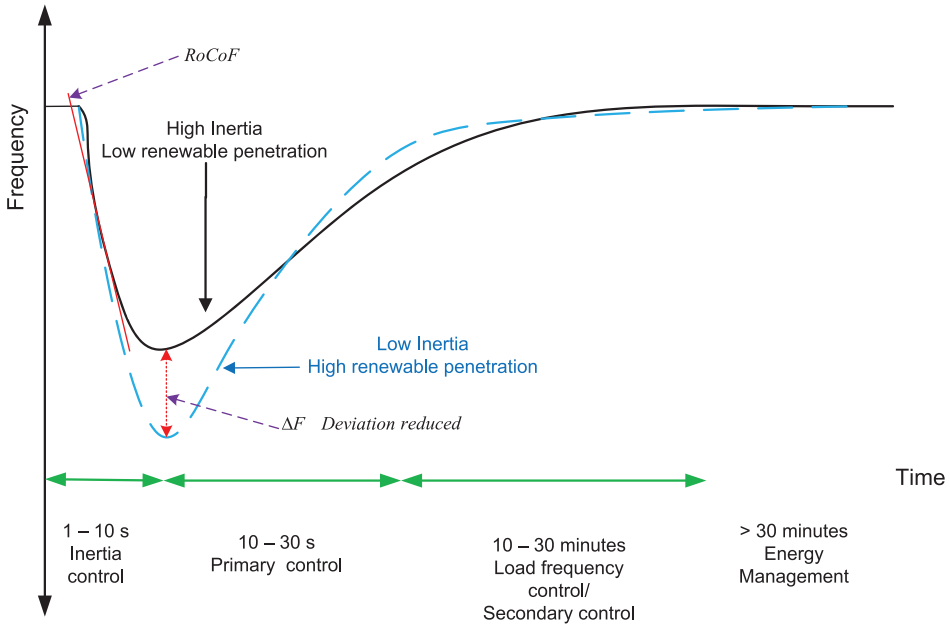


FIGURE 6 Main frequency deviation curve.

as follows:

$$w = (w_{start} - w_{end}) \times \left(\frac{T_{\max} - t}{T_{\max}}\right) + w_{end} \times e^{-t/\left(\frac{T_{\max}}{10}\right)} \quad (12)$$

where $w_{start} = 0.9$ and $w_{end} = 0.2$

The flowchart in Figure 7 is an algorithm for MPSO, modified inertia weight as inserted in Equations (10) and (12).

It is worth mentioning that the power required to maintain the frequency stability of the system may be determined by analysing the discrepancy between the voltage and normalized frequency. For adjustment and better frequency stability support, the PID controller is applied for correction in the differences [41], as stated in Equation (13), which is an extension of Equation (7), illustrating the achievement of reference active power.

$$P_{ref} = \left(\frac{\Delta f}{f_{nom}} - \frac{\Delta V}{V_{nom}}\right) \cdot \left(K_{P1} + \frac{K_{I1}}{S} + K_{D1}S\right) \quad (13)$$

In this proposed control strategy, power flows out from the ESS when there is increase in load or when a supply shuts down. When RES source is added, the power flows to the ESS. The converter acts as bidirectional virtual support. Virtual inertia control is critical for frequency stability as it reduces RoCoF and frequency zenith and nadir.

3.2 | Modelling of block diagrams of frequency control

The frequency control block is shown in the diagram, the frequency Q_g, f_g, and P_g are the measured grid reactive power, frequency, and real power respectively. The outcomes of the frequency control are displacement angle θ_o, output frequency f_o, and angular frequency ω_o. The controller operates in a similar manner as standard frequency controller of synchronous generators [42]. Figure 8 shows switching signals which control the gates. The frequency control diagram is shown in Figure 9.

4 | ANALYSIS OF SIMULATION AND EXPERIMENTAL FINDINGS AND DISCUSSION

4.1 | Simulation findings and discussion

The effectiveness of the proposed technique was evaluated by simulating the frequency of the microgrid using Matlab/Simulink. The ultracapacitor is designed to improve frequency responsiveness. Parameter tuning is done by the use of optimization techniques. As shown, frequency deviation Δf is the optimal objective function,

$$ISE = \int_0^{T_{Sim}} |\Delta f^2| dt \quad (14)$$

where

ISE is integral squired error
T_{Sim} is simulation time range

The objective function in Equation (14) is applied using both optimization techniques and the outcomes are recorded in Table 2. In performance index ISE, the error is squared; therefore, it handles both positive and negative errors.

The cases considered in the analysis of this proposed controller in the MG's frequency response are as follows:

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pel2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

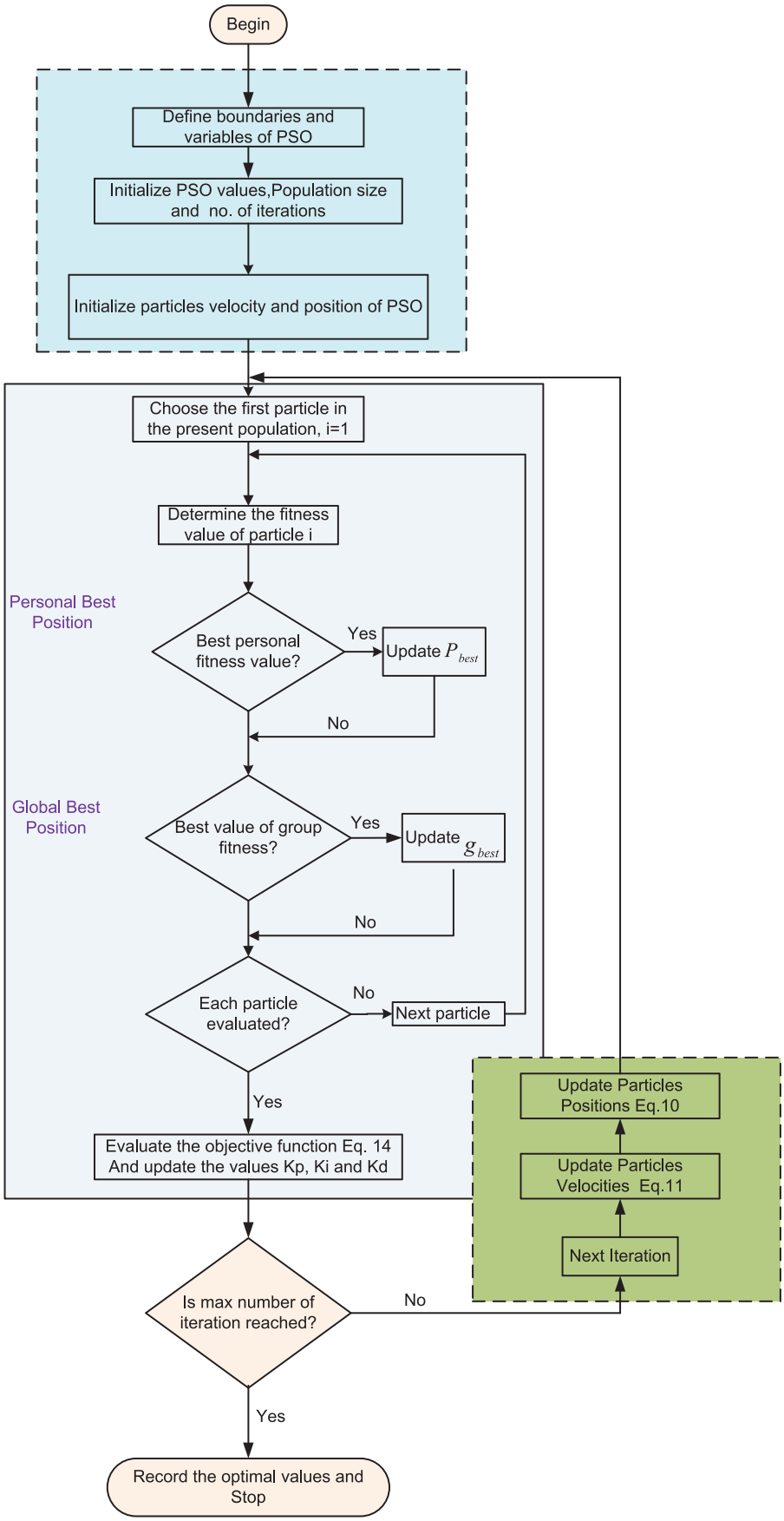


FIGURE 7 Improved particle swarm optimization (MPSO) flowchart.

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pel2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

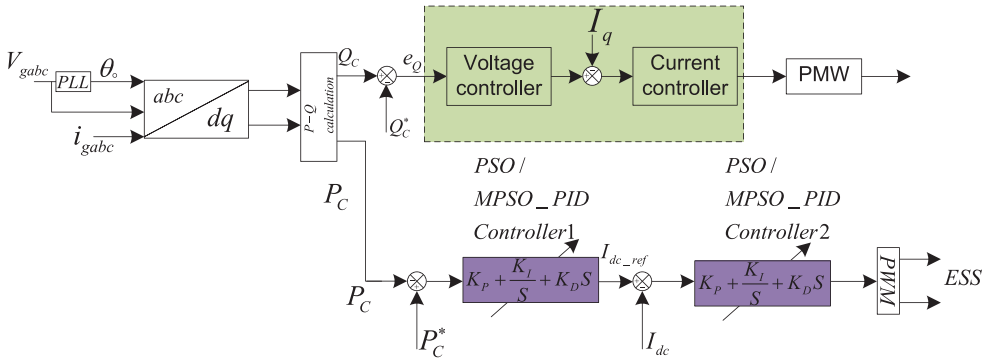


FIGURE 8 Block diagrams of control frame for switching signals.

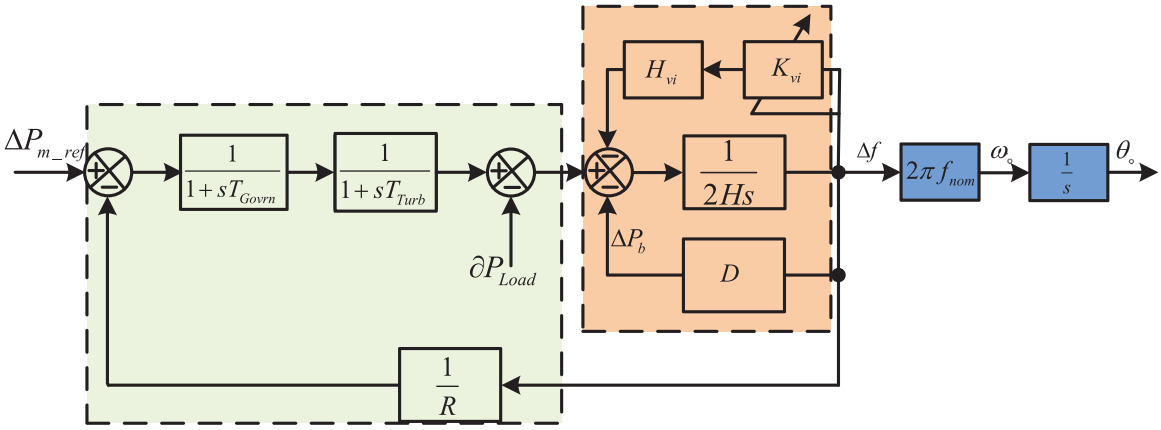


FIGURE 9 Block diagram of inertia control structure of frequency regulation.

TABLE 2 Optimal values.

Parameters	PSO	Improved PSO
K_P	1000	1000
K_I	39.0995	0.6893
K_D	0	0

Abbreviation: PSO, particle swarm optimization.

CASE 1: Increased power generation of 5 MW from PV

Figure 10a illustrates the frequency response of microgrid due sudden increase in PV generated power at 0.2 s by virtue of increased radiations. The frequency rose abruptly post disturbance attributed to power generated by PV unit. Both the convectional PSO and Improved PSO have been implemented for parameters tuning in such a way that the deviation in frequency is contained within the permissible threshold. However, Improved PSO has smaller frequency deviation compared with conventional PSO; that is, 50. 0199 Hz and 50.997, respectively. The difference in frequencies is used to calculate the capacity of ultracapacitor energy saved by applying Equation (2). The difference in frequencies using both the methods is found to be 0.98 Hz which is equivalent to additional amount of energy storage of 490 kW.s. needed to curtail the frequency deviation. In other words, the improvised PSO helps in reducing the additional storage capacity in comparison to conventional PSO.

CASE 2: Increase in load demand of 7.5 MW

In this scenario, a sudden surge in load demand of 7.5 MW is considered at 3 s. Due to this disturbance, frequency drop is seen just after the increase in load as depicted in Figure 11a. In this scenario, the lowest point is 48.52 Hz using conventional PSO, though it is within allowable frequency deviation but it requires large ultracapacitor to accommodate the fluctuation of such magnitude. For improved PSO, the frequency deviation is found to be 49.97 Hz. The difference between the two strategies in terms of frequency is 1.45 Hz. Thus, by applying Equation (2), the additional storage capacity of 1.0875 MW.s is saved using proposed Improvised PSO. For the convenience of comparison, the disturbance for scenario 1 was kept constant throughout the simulation at 0.2 of 5 MW.s. Furthermore, the fact that this technique can handle both circumstances at once is proof that it is better.

CASE 3: Increase in wind generated power of 10 MW

In this scenario, a sudden enhanced wind power of 10 MW is considered at 6 s due to sudden increase in wind velocity. Due to this sudden increase, imbalance between demand and supply arises which results in frequency shoot up as shown in Figure 12a. The highest point is 51.9886 Hz based on PSO and 50.0398 Hz based on Improved PSO. It is to be noticed that the

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pel2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

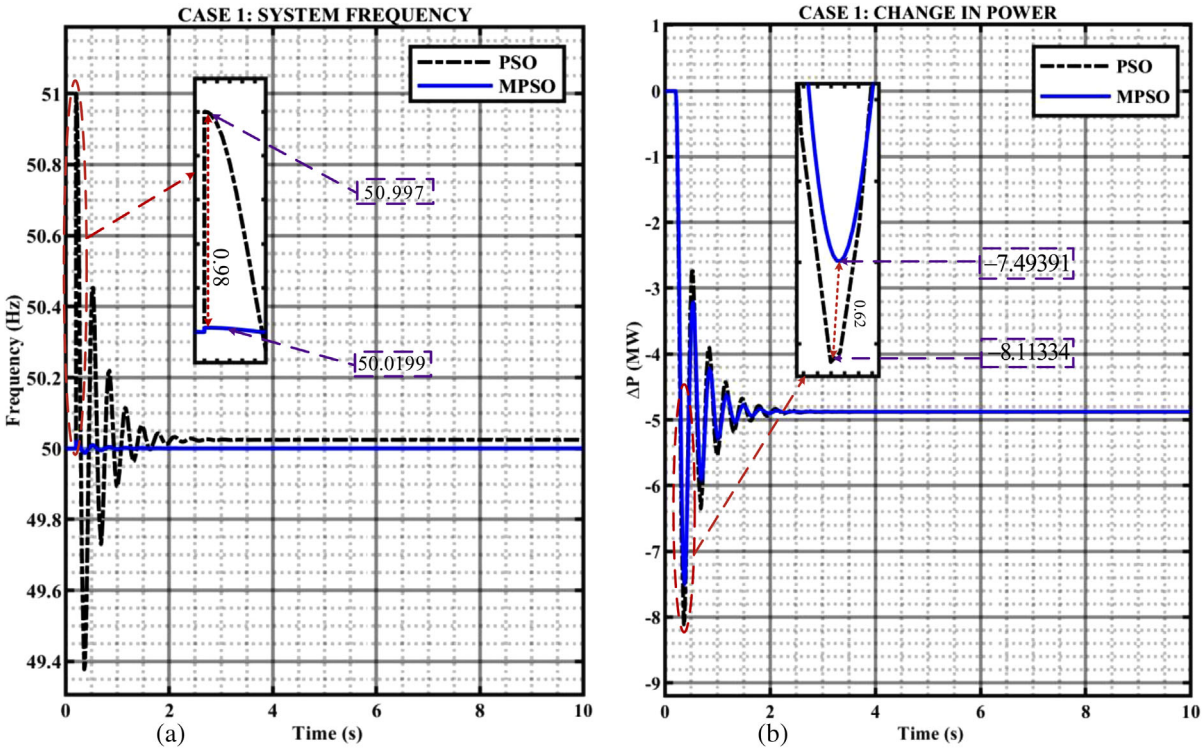


FIGURE 10 (a) Frequency response, (b) 5 MW increased power generation of 5 MW from PV.

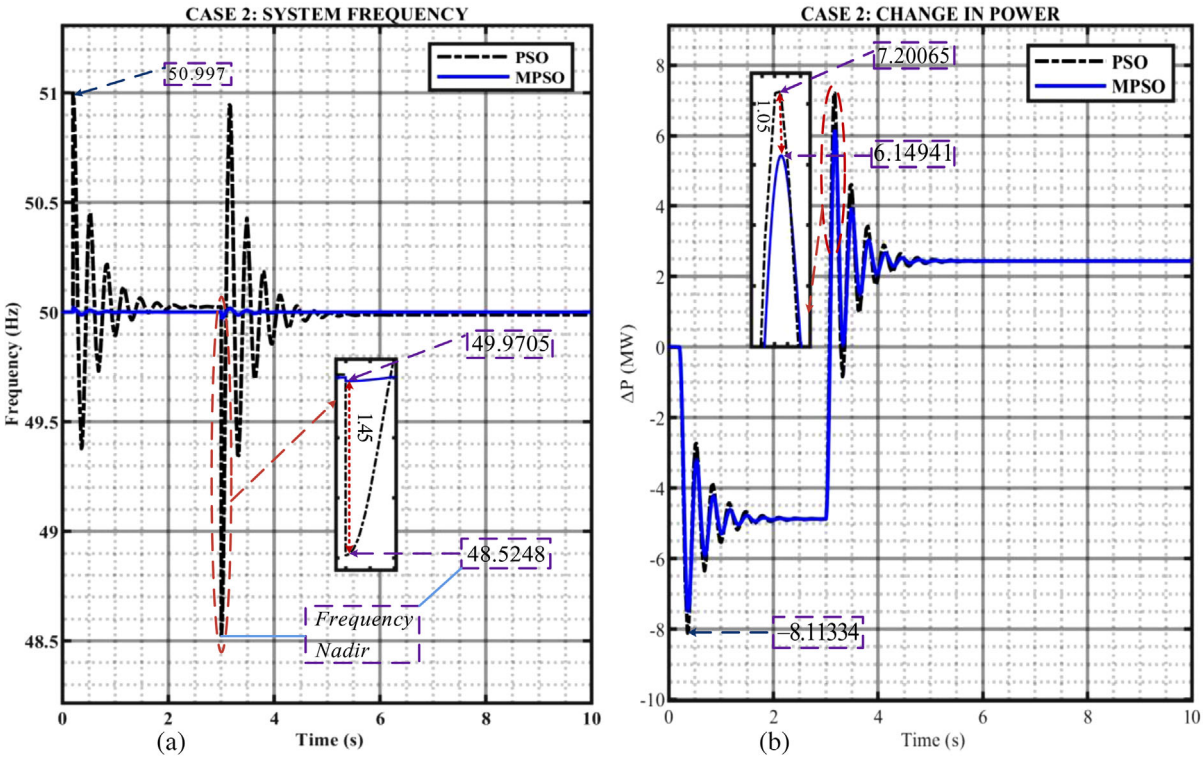


FIGURE 11 (a) Frequency response. (b) Increase in load demand of 7.5 MW.

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pe2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

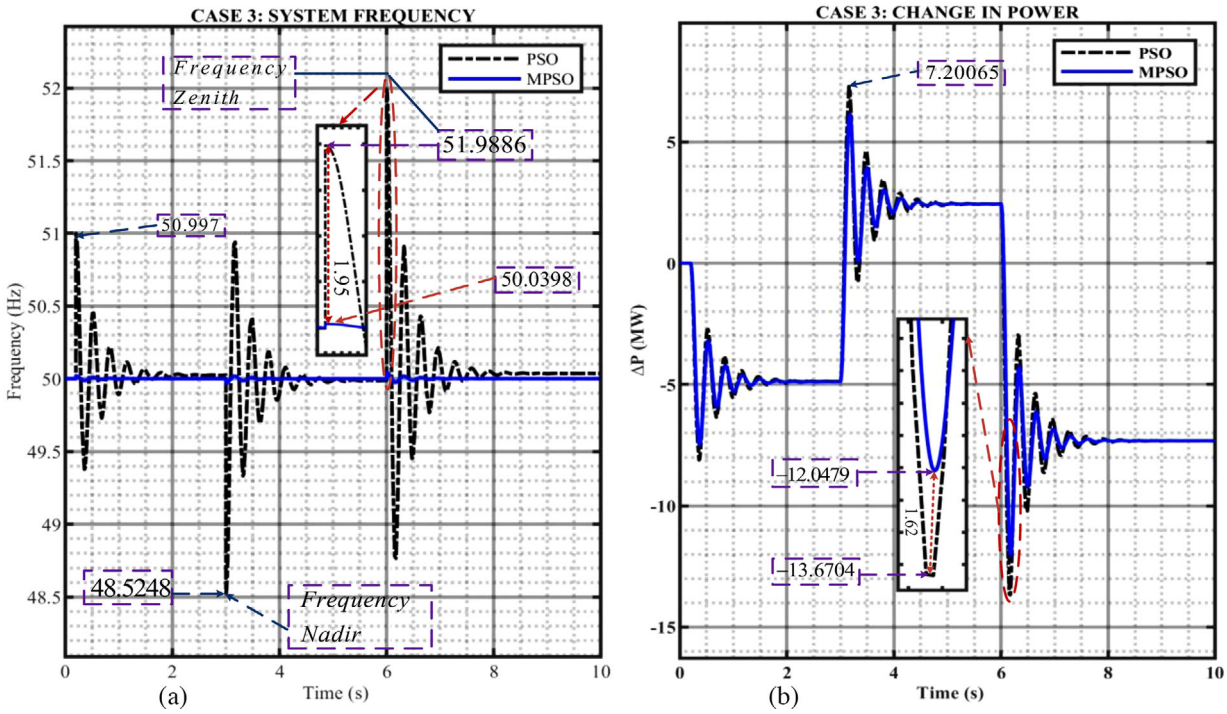


FIGURE 12 (a) Frequency response. (b) Increase in wind generated power of 10 MW.

TABLE 3 Microgrid frequency deviations.

	Frequency deviation (Hz)	
Case	**Convectional PSO**	**Improved PSO**
1	0.997	0.0199
2	1.4752	0.0295
3	1.9886	0.0398

Abbreviation: PSO, particle swarm optimization.

differences in frequency using both the strategies are 1.95 Hz. The energy calculated for this much frequency difference is of 1.95 MW.s. This shows that the cost of installing larger ultracapacitor is reduced with Improved PSO. To make comparisons simple, disturbances for Case 1 at 0.2 s of 5 MW and Case 2 at 3 s of 7.5 MW were included during simulation. Given that it can manage all three eventualities at once, this technique provides further proof of its superiority. The results of the simulations show that frequency is restored in every instance.

From Table 3, are tabulated results of the simulation of the microgrid frequency deviation. It was conducted in all test conditions. The findings clearly demonstrate that the proposed controller outperformed the convectional method. The proposed method is ensuring frequency stabilization and economical use of ultracapacitor. Here, it could be observed that for a minimum disturbance of 5 MW, the change in frequency is

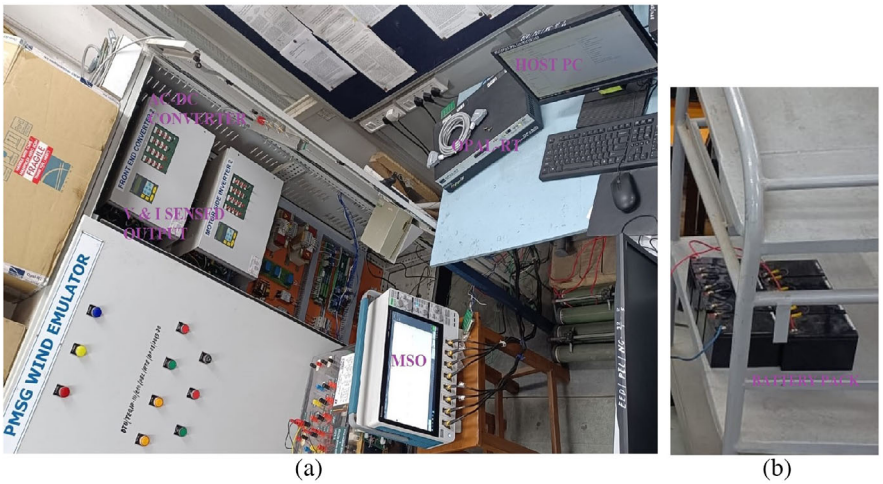


FIGURE 13 Photo of experimental setting. (a) Setup. (b) Battery pack.

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pel2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 Sum up of the frequency response under various severe disturbances.

	Disturbance	Frequency response
Load fluctuation	Load increasing	Decrease (nadir)
	Load shedding (decreasing)	Increase (zenith)
Wind speed variation	Speed increment	Increase (zenith)
	Speed decrement	Decrease (nadir)
Solar irradiation variation	Radiations increment	Increase (zenith)
	Radiations decrement	Decrease (nadir)
Contingency	RES disconnecting	Decrease (nadir)
	RES connecting	Increase (zenith)

0.997 (1.994 %) while with Improved PSO frequency change is 0.0199 (0.0398%). Similarly with a maximum disturbance of 10 MW, the change in frequency is 1.9886 (3.977%) while Improved PSO yields a frequency change of 0.0398 (0.0796%). Thus, it is evident from the two extreme cases that with Improved PSO the change in frequency is very small. For brevity purposes, Table 4 is a summary of frequency response in a microgrid under different disturbance scenarios.

4.2 | Experimental results and discussion

The simulation study carried out in the MATLAB/Simulink environment is validated in real time using (OP-4510) OPAL-RT as demonstrated in Figure 13. This offers as further evidence of the effectiveness of the proposed study concept in a real-time setting.

The fast dynamic response of the proposed dynamic controller is apparent; it is capable of quickly reducing the frequency deviation in under 2 s, regardless of the degree of the disturbance. In the experimental study, disturbances of 2%, 3%, and 4% were introduced and it can be observed that the response times in all three situations a found to be very quick. The

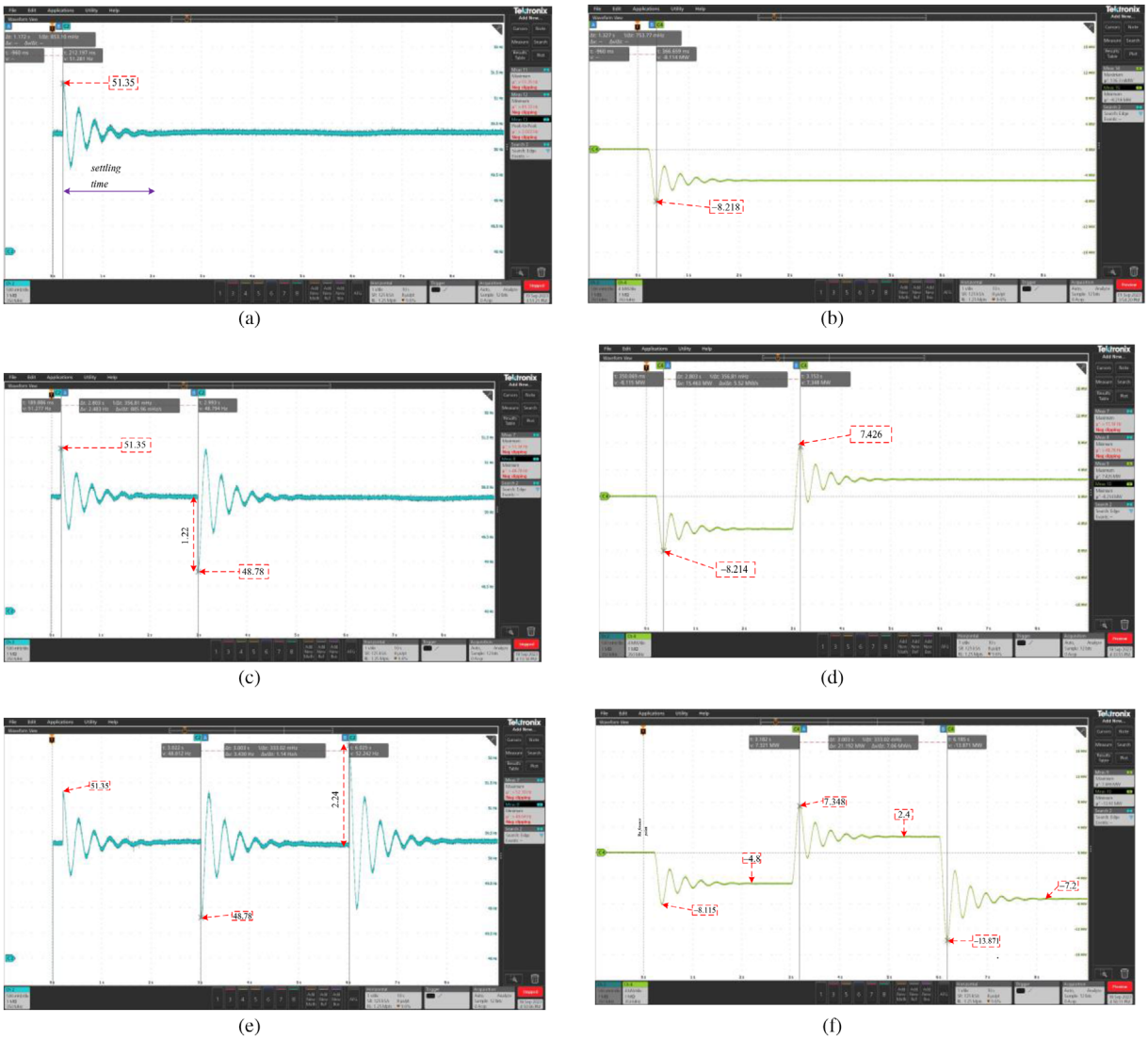


FIGURE 14 Frequency deviation (a) 5 MW, (c) 7.5 MW, (e) 10 MW. Power change (b) 5 MW, (d) 7.5 MW, (f) 10 MW.

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pel2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5 Distinction of proposed control technique over other methods.

Reference	Type of a controller	Contribution
[43]	Hybrid ESS based on heuristics methods	PID controller parameters are tuned using GA and PSO but suffers local optima problem
[29]	Optimal virtual synchronous generator control of HESS	Designing optimal VSG using PSO but suffers local optima and frequency oscillations take long to settle
[44]	PID controller based on the Grey Wolf Optimization (GWO) algorithm	Frequency regulation in the microgrid designed for a multi-source single-area system; however, it does not take into account multiple areas or the impact of (RES).
[45]	BFOA tuned LQR-based VSG	Suppress frequency variations using robust design strategies. It suffers convergence challenge.
Proposed	HESS based on improved PSO	Optimizing PID controller settings using an enhanced version of PSO and traditional PSO. The Improved PSO algorithm is online optimization technique which successfully addressed the difficulties of local optima and lengthen iterations.

Abbreviations: BFOA, bacteria foraging optimization algorithm; HESS, hybrid energy storage system; LQR, linear-quadratic regulator; PID, proportional (Kp), integral (Ki), and derivative (Kd) controller; PSO, particle swarm optimization; VSG, virtual synchronous generators.

results from the real-time OPAL-RT (OP-4510) and MATLAB simulations are consistent, indicating that offline systems may employ the adjusted optimum PID controller parameters. This is done to support the suggested control strategy, which virtually emulates simulation results with hardware in loop.

As seen in Figures 14a to f, the same three cases have been implemented using HIL. It is worthwhile noting that with disturbance in load operating conditions, the disturbance in frequency is restored as per grid code requirements. Moreover, the restoration period using the proposed method is found to be very much smaller in comparison to conventional methods which further results in an enhanced stability.

Finally, a brief comparison and distinction with the existing state-of-the-art methods of control in grid-connected ESSs have been provided in Table 5.

5 | CONCLUSION

In this paper, frequency stabilization of RESs based low inertia microgrid has been achieved with HESS. The optimal sizing of HESS is done in such a way that it utilizes the response characteristics of Ultracapacitor and battery in order to have enhanced performance with reduced cost. Ultracapacitor is used to mimic inertia for high frequency power deviation as seen during simulation, whereas battery is applied for long-term power variation in microgrid. A modified form of particle swarm optimization was used to optimize the PID factors. The efficiency of the proposed control strategy has been proved by examining three distinct levels of uncertainty generated by either intermittent RESs or changing load. There has been a considerable decrease in the problems related to the rate of change of frequency, overshoot, undershoot, and rising time. As a result, the required size of ultracapacitors has been reduced. In turn, the cost of the entire system is reduced. The effectiveness of modified particle swarm optimization is evident when comparing it to traditional particle swarm optimization. It is worthy to note that there is significant amount of energy saving capacity of 1. 95 MW.s for a disturbance of 10 MW. Moreover, the conventional particle swarm optimization technique yielded a frequency deviation of 1.9886 Hz, but the proposed Improved particle swarm optimization technique successfully decreased it to 0.0398 Hz. Therefore, the suggested method enhances the overall efficiency of the microgrid across various operational scenarios. The simulation findings, together with the experimental findings, confirm the efficacy of the proposed strategy in terms of determining the appropriate size of the Hybrid Energy Storage System (HESS) and enhancing the control performance of the Microgrid. The future investigation should consider seamless power sharing in HESS and the ratios based on online optimization.

AUTHOR CONTRIBUTIONS

Philemon Yegon: Conceptualization; data curation; formal analysis; investigation; methodology; software. **Mukhtiar Singh**: Supervision; validation; visualization; writing—review & editing.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

FUNDING INFORMATION

No funding has been received from any source for the implementation of this research. This research work is entirely the product of authors' initiatives.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon request.

ORCID

Philemon Yegon https://orcid.org/0009-0003-3179-1511

REFERENCES

1. Fang, J., Li, H., Tang, Y., Blaabjerg, F.: On the inertia of future more-electronics power systems. IEEE J. Emerg. Sel. Top. Power Electron. 7(4), 2130–2146 (2019). https://doi.org/10.1109/JESTPE.2018.2877766
2. Chang-Chien, L.R., An, L.N., Lin, T.W., Lee, W.J.: Incorporating demand response with spinning reserve to realize an adaptive frequency restoration plan for system contingencies. IEEE Trans. Smart Grid 3(3), 1145–1153 (2012). https://doi.org/10.1109/TSG.2012.2192297
3. Liu, K., Qu, Y., Kim, H.M., Song, H.: Avoiding frequency second dip in power unreserved control during wind power rotational speed recovery. IEEE Trans. Power Syst. 33(3), 3097–3106 (2018). https://doi.org/10.1109/TPWRS.2017.2761897
4. ISO 2013: Reciprocating internal combustion engine driven alternating current generating sets — Part 5: Generating sets. 2013. https://cdn.standards.iteh.ai/samples/81844/3dd696ded80349f29547ba03decd6705/ISO-8528-5-2022.pdf

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pe2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5. NERC: Frequency Response Initiative Report. 2012. https://www.nerc.com/pa/Stand/Project200712FrequencyResponseDL/Bal-003-1-Background_Document-Clean-2013_FILING.pdf
6. Huang, L., Xin, H., Wang, Z.: Damping low-frequency oscillations through VSC-HVdc stations operated as virtual synchronous machines. IEEE Trans. Power Electron. 34(6), 5803–5818 (2019). https://doi.org/10.1109/TPEL.2018.2866523
7. Soni, N., Doolla, S., Chandorkar, M.C.: Inertia design methods for islanded microgrids having static and rotating energy sources. IEEE Trans. Ind. Appl. 52(6), 5165–5174 (2016). https://doi.org/10.1109/TIA.2016.2597281
8. Yuan, C., Yang, D., Feng, J., Liu, C.: Constrained operation zone of a VSG considering the DC-side power margin. J. Eng. 2019(16), 2563–2568 (2019). https://doi.org/10.1049/joe.2018.8679
9. Meng, J., et al.: An overview and comparison of online implementable SOC estimation methods for lithium-ion battery. IEEE Trans. Ind. Appl. 54(2), 1583–1591 (2018). https://doi.org/10.1109/TIA.2017.2775179
10. Ma, Y., Cao, W., Yang, L., Wang, F.F., Tolbert, L.M.: Virtual synchronous generator control of full converter wind turbines with short-term energy storage. IEEE Trans. Ind. Electron. 64(11), 8821–8831 (2017). https://doi.org/10.1109/TIE.2017.2694347
11. Sun, C., Ali, S.Q., Joos, G., Bouffard, F.: Design of hybrid-storage-based virtual synchronous machine with energy recovery control considering energy consumed in inertial and damping support. IEEE Trans. Power Electron. 37(3), 2648–2666 (2022). https://doi.org/10.1109/TPEL.2021.3111482
12. Chapaloglou, S., Alves, E.F., Trovato, V., Tedeschi, E.: Optimal energy management in autonomous power systems with probabilistic security constraints and adaptive frequency control. IEEE Trans. Power Syst. 39(1), 1543–1554 (2024). https://doi.org/10.1109/TPWRS.2023.3236378
13. Bhattar, C.L., Chaudhari, M.A.: Centralized energy management scheme for grid connected DC microgrid. IEEE Syst. J. 17(3), 3741–3751 (2023). https://doi.org/10.1109/JSYST.2022.3231898
14. Ikram, A.I., Ullah, A., Datta, D., Islam, A., Ahmed, T.: Optimizing energy consumption in smart homes: Load scheduling approaches. IET Power Electron. 2023, 1–13 (2024). https://doi.org/10.1049/pel2.12663
15. Chen, Y., Meng, J., Yan, Y., Zhong, L.: Quasi-Z based adaptive sliding mode control for three-phase photovoltaic grid-connected system. IET Power Electron. 16(15), 2577–2591 (2023). https://doi.org/10.1049/pel2.12584
16. Majeed, M.A., Phichisawat, S., Asghar, F., Hussan, U.: Optimal energy management system for grid-tied microgrid: An improved adaptive genetic algorithm. IEEE Access 11, 117351–117361 (2023). https://doi.org/10.1109/ACCESS.2023.3326505
17. Joshi, M.C., Samanta, S.: Energy management with improved frequency sharing based control for battery/ultracapacitor hybrid energy system in the presence of delay. IET Power Electron. 13(10), 2019–2028 (2020). https://doi.org/10.1049/iet-pel.2018.5118
18. Maleki, S., Nikoukar, J., Tousifian, M.H.: Robust frequency control of microgrids: A mixed H2/H∞ virtual inertia emulation. Int. Trans. Electr. Energy Syst. 2023, 1 (2023). https://doi.org/10.1155/2023/6872765
19. Golpîra, H., Seifi, H., Messina, A.R., Haghifam, M.R.: Maximum penetration level of micro-grids in large-scale power systems: Frequency stability viewpoint. IEEE Trans. Power Syst. 31(6), 5163–5171 (2016). https://doi.org/10.1109/TPWRS.2016.2538083
20. Guo, L., Xu, Z., Jin, N., Li, Y., Wang, W.: A weighted voltage model predictive control method for a virtual synchronous generator with enhanced parameter robustness. Prot. Control Mod. Power Syst. 6(1), 1–11 (2021). https://doi.org/10.1186/s41601-021-00217-8
21. Elwakil, M.M., Zoghaby, H.M.E., Sharaf, S.M., Mosa, M.A.: Adaptive virtual synchronous generator control using optimized bang-bang for Islanded microgrid stability improvement. Prot. Control Mod. Power Syst. 8(1), 57 (2023). https://doi.org/10.1186/s41601-023-00333-7
22. Shi, T., Sun, J., Han, X., Tang, C.: Research on adaptive optimal control strategy of virtual synchronous generator inertia and damping parameters. IET Power Electron 17(1), 121–133 (2024). https://doi.org/10.1049/pel2.12620
23. Skiparev, V., et al.: Virtual inertia control of isolated microgrids using an NN-based VFOPID controller. IEEE Trans. Sustain. Energy 14(3), 1558–1568 (2023). https://doi.org/10.1109/TSTE.2023.3237922
24. Leng, D., Polmai, S.: Virtual synchronous generator based on hybrid energy storage system for PV power fluctuation mitigation. Appl. Sci. 9(23), 1–18 (2019). https://doi.org/10.3390/app9235099
25. Li, D., Zhu, Q., Lin, S., Bian, X.Y.: A self-adaptive inertia and damping combination control of VSG to support frequency stability. IEEE Trans. Energy Convers. 32(1), 397–398 (2017). https://doi.org/10.1109/TEC.2016.2623982
26. Kerdphol, T., Watanabe, M., Hongesombut, K., Mitani, Y.: Self-adaptive virtual inertia control-based fuzzy logic to improve frequency stability of microgrid with high renewable penetration. IEEE Access 7, 76071–76083 (2019). https://doi.org/10.1109/ACCESS.2019.2920886
27. Hu, Y., Wei, W., Peng, Y., Lei, J.: Fuzzy virtual inertia control for virtual synchronous generator. In: Chinese Control Conference CCC 2016, Aug., pp. 8523–8527 (2016). https://doi.org/10.1109/ChiCC.2016.7554718
28. Alipoor, J., Miura, Y., Ise, T.: Power system stabilization using virtual synchronous generator with alternating moment of inertia. IEEE J. Emerg. Sel. Top. Power Electron. 3(2), 451–458 (2015). https://doi.org/10.1109/JESTPE.2014.2362530
29. Mohamed, M.M., El Zoghby, H.M., Sharaf, S.M., Mosa, M.A.: Optimal virtual synchronous generator control of battery/supercapacitor hybrid energy storage system for frequency response enhancement of photovoltaic/diesel microgrid. J. Energy Storage 51, 104317 (2022). https://doi.org/10.1016/j.est.2022.104317
30. Chong, L.W., Wong, Y.W., Rajkumar, R.K., Isa, D.: An optimal control strategy for standalone PV system with battery-supercapacitor hybrid energy storage system. J. Power Sources 331, 553–565 (2016). https://doi.org/10.1016/j.jpowsour.2016.09.061
31. Le Nguyen, V.T.N.T.B.: Stability analysis of an isolated microgrid with the presence of the hybrid energy storage system-based virtual synchronous generator. J. Sci. Technol. Issue Inf. Commun. Technol. 46–51 (2020). https://doi.org/10.31130/jst-ud2020-101e
32. Li, K., Cheng, P., Wang, L., Tian, X., Ma, J., Jia, L.: Improved active power control of virtual synchronous generator for enhancing transient stability. IET Power Electron. 16(1), 157–167 (2023). https://doi.org/10.1049/pel2.12371
33. Oh, Y.G., Kwon, O., Kwon, J.M.: Bidirectional push–pull/H-bridge converter for low-voltage energy storage system. IET Power Electron. 17(1), 1–9 (2024). https://doi.org/10.1049/pel2.12586
34. Saadat, H. Power System Analysis. McGraw Hills-COMPANIES, New York (1999). https://doi.org/10.1201/9781003394433-4
35. Fang, J., Li, H., Tang, Y., Blaabjerg, F.: Distributed power system virtual inertia implemented by grid-connected power converters. IEEE Trans. Power Electron. 33(10), 8488–8499 (2018). https://doi.org/10.1109/TPEL.2017.2785218
36. James Kennedy, R.E.: Particle swarm optimization. In: Proceedings of ICNN'95 - International Conference on Neural Networks, Perth, WA, Australia, pp. 1942–1948 (1995). https://doi.org/10.1109/ICNN.1995.488968
37. Arasomwan, M.A., Adewumi, A.O.: On the performance of linear decreasing inertia weight particle swarm optimization for global optimization. Sci. World J. 2013, 1–12 (2013). https://doi.org/10.1155/2013/860289
38. Yegon, P., Singh, M.: Frequency stability enhancement of microgrid using optimization techniques-based adaptive virtual inertia control. Int. Trans. Electr. Energy Syst. 2023, 1 (2023). https://doi.org/10.1155/2023/2121721
39. Oladipo, S., Sun, Y., Adeleke, O.: An improved particle swarm optimization and adaptive neuro-fuzzy inference system for predicting the energy consumption of university residence. Int. Trans. Electr. Energy Syst. 2023, 1 (2023). https://doi.org/10.1155/2023/8508800
40. Guimin, C., Xinbo, H., Jianyuan, J., Zhengfeng, M.: Natural exponential inertia weight strategy in particle swarm optimization. Proc. World Congr. Intell. Control Autom. 1(2002), 3672–3675 (2006). https://doi.org/10.1109/WCICA.2006.1713055
41. Zhang, Z., Fang, J., Dong, C., Jin, C., Tang, Y.: Enhanced grid frequency and DC-link voltage regulation in hybrid AC/DC microgrids through

17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pel2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

bidirectional virtual inertia support. IEEE Trans. Ind. Electron. 70(7), 6931–6940 (2023). https://doi.org/10.1109/TIE.2022.3203757
42. Kundur, P. Power System Stability and Control. McGraw Hill Education (2007). https://doi.org/10.1201/9781420009248
43. Guentri, H., Allaoui, T., Mekki, M., Denai, M.: Power management and control of a photovoltaic system with hybrid battery-supercapacitor energy storage based on heuristics methods. J. Energy Storage 39, 102578 (2021). https://doi.org/10.1016/j.est.2021.102578
44. Paliwal, N., Srivastava, L., Pandit, M.: Application of grey wolf optimization algorithm for load frequency control in multi-source single area power system. Evol. Intell. 15(1), 563–584 (2022). https://doi.org/10.1007/s12065-020-00530-5
45. Singh, S.K., Singh, R., Ashfaq, H., Kumar, R.: Virtual inertia emulation of inverter interfaced distributed generation (IIDG) for dynamic frequency stability & damping enhancement through BFOA tuned optimal controller. Arab. J. Sci. Eng. 47(3), 3293–3310 (2022). https://doi.org/10.1007/s13369-021-06121-5

How to cite this article: Yegon, P., Singh, M.: Optimization of battery/ultra-capacitor hybrid energy storage system for frequency response support in low-inertia microgrid. IET Power Electron. 1–15 (2024). https://doi.org/10.1049/pel2.12723

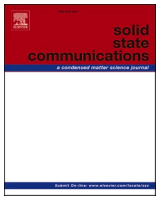
17554543, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/pel2.12723 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Solid State Communications 389 (2024) 115590

Contents lists available at ScienceDirect

Solid State Communications

journal homepage: www.elsevier.com/locate/ssc



Performance analysis of $LaFeO_3$ perovskite solar cells: A theoretical and experimental study

Rahul Kundara *, Sarita Baghel

Department of Applied Physics, Delhi Technological University, Delhi, India

ARTICLE INFO

Communicated by: Francois Peeters

Keywords:
$LaFeO_3$ perovskite material
Sol-gel method
SCAPS-1D
ETLs and HTLs

ABSTRACT

Lanthanum ferrite ($LaFeO_3$) is a potential candidate for Perovskite Solar Cell (PSC) application. The transition metal oxide has strong electron-electron correlation which results in an increase in quantum efficiency (QE) of PSC. In this investigation, we reported synthesis of $LaFeO_3$ perovskite material, simulation of $LaFeO_3$-based PSC followed by experimental validation. The sol-gel method was employed to prepare nanoparticles of $LaFeO_3$ for PSC application. The prepared perovskite material is analyzed by XRD, SEM, EDS, UV–Vis and FTIR. The photovoltaic (PV) application of synthesized $LaFeO_3$ material is studied by SCAPS-1D. The $LaFeO_3$-based PSC is simulated with various ETLs such as TiO_2, PCBM, IGZO, In_2O_3, C_{60}, ZnO, SnO_2, WS_2 and HTLs of Cu_2O, $CuInSe_2$, Spiro-OMeTAD, P3HT, D-PBTTT-14, V_2O_5, CFTS, CBTS. The simulation outcomes of 64 distinct configurations deduced the eight-best performing PSCs. For these eight PSCs, impact of thickness of perovskite absorber layer, operating temperature, defect density, series and shunt resistance, work function of back contact, J-V characteristics and QE curve has been studied. We have achieved theoretical maximum efficiency of 13.03 % for the configuration of FTO/WS_2/$LaFeO_3$/CBTS/Au. The optimized structure has experimentally obtained maximum efficiency of 0.79 %. This shows that it is promising material for future PV application.

1. Introduction

The limited amount of fossil fuels and their combustion, leads to pollution in the environment [1]. So, alternative energy technologies are required like biomass, wind, solar and tidal power. Solar energy is a green, accessible and abundant renewable energy [2–4]. Perovskite solar cell (PSC) continues to draw the attention of researchers by virtue of characteristics like high-power conversion efficiency (PCE) and simple room temperature fabrication methodology which makes them a cheaper alternative to conventional silicon solar cells [5–7]. A perovskite has basically ABX_3 type structure (where A: alkyl group, B: metal and X: halogen) [8,9]. Pb-based PSCs present sustainability challenges due to high toxicity of lead, posing risks to both living organisms and the ecosystem. Moreover, Pb-based PSCs suffer from stability issues stemming from the rapid oxidation of Pb cation. In the quest for a suitable Pb substitute [10], researchers have explored numerous candidates to replace Pb, such as Sn [11], Ge [12], Bi [13], Sb [14], La [15,16] and double perovskites etc. $LaFeO_3$ possesses a tunable energy band-gap which can be altered using suitable dopants [17,18]. The other properties of $LaFeO_3$ such as chemical stability, low production cost, high electron mobility and its non-toxic nature are already favorable for lead-free PSC application [19,20]. There are very few studies on $LaFeO_3$ as perovskite material for solar cell application [21]. This work comprises fabrication and characterization of $LaFeO_3$-based PSCs. $LaFeO_3$ nanoparticles were first synthesized by using the sol-gel method and then they were used for solar cell application. Prior to fabrication of $LaFeO_3$-based PSC, we have simulated and optimized the structure of solar cells using SCAPS-1D simulation software. So, we have got eight optimized structures of $LaFeO_3$-based PSC through simulation with a maximum PCE of 13.03 % for the most efficient configuration which is approximately 30 % greater than the past reported work [20]. The experimental validation was conducted for the optimized PSC. We achieved an PCE of 0.79 % experimentally, which is approximately thirteen times higher than the previously reported result [21]. The spin coating method is employed to obtain PCE of 0.79 % which is the highest achieved efficiency for the $LaFeO_3$-based PSC till now. This efficiency enhancement is attributed to proper selection of hole transport layer (HTL) and electron transport layer (ETL). A conventional PSC structure contains a perovskite absorber layer sandwiched amid ETL and HTL. A current is developed when electrons (e^-) are traveled through ETL and

* Corresponding author.
E-mail address: rahul_2k20phdap10@dtu.ac.in (R. Kundara).

https://doi.org/10.1016/j.ssc.2024.115590
Received 24 January 2024; Received in revised form 20 May 2024; Accepted 12 June 2024
Available online 12 June 2024
0038-1098/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

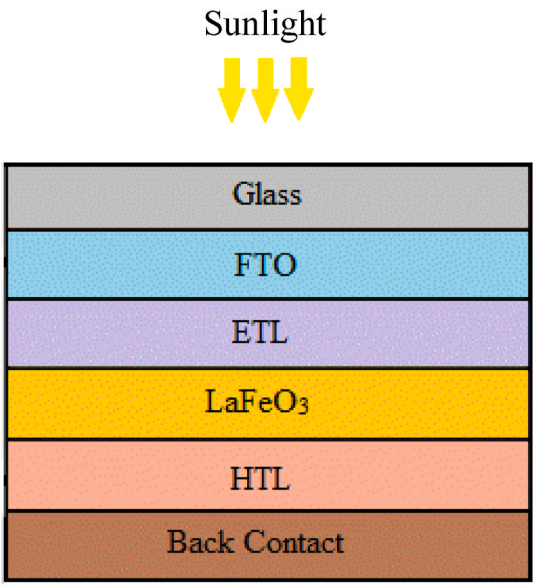


Fig. 1. Architecture of $LaFeO_3$-based PSC.

holes (h^+) are transferred to HTL from the absorber layer. To obtain the maximum PCE of a PSC, the energy level alignment of all these layers becomes a very important factor. The selection and optimization of the ETL and HTL is very crucial to maximize the performance of a PSC. In this simulation we have simulated sixty-four configurations with different ETLs (TiO_2, PCBM, IGZO, In_2O_3, C_{60}, ZnO, SnO_2, WS_2) and HTLs (Cu_2O, $CuInSe_2$, Spiro-OMeTAD, P3HT, D-PBTTT-14, V_2O_5, CFTS, CBTS). The outcomes showed that CBTS is the most suitable HTL for $LaFeO_3$ PSC and in addition it is abundantly available [22]. The study also includes optimization of the parameters which influences performance of the PSC, thickness, operating temperature, defect density (N_t), back contact, series (R_S) and shunt resistance (R_{SH}).

2. Device architecture

The PSC structure consists of ETL, absorber layer and HTL shown below in Fig. 1. The $LaFeO_3$ is employed as an absorber layer in PSC having thickness of 400 nm. In this simulation, we have employed several ETLs (TiO_2, PCBM, IGZO, In_2O_3, C_{60}, ZnO, SnO_2, WS_2) and HTLs (Cu_2O, $CuInSe_2$, Spiro-OMeTAD, P3HT, D-PBTTT-14, V_2O_5, CFTS, CBTS) to achieve the maximum efficiency. The thickness of employed ETL = 100 nm and HTL = 50 nm. In p-i-n devices, the electrons and holes generated within the absorber layer when sunlight falls on it. The generated e^- and h^+ is traveled through ETL and HTL, e^-s are collected at the fluorine doped tin oxide (FTO) and h^+s at the back contact.

3. Experimental section

3.1. Materials and method

All the precursors employed in synthesis of perovskite are $La(NO_3)_3$, $Fe(NO_3)_3$, citric acid and ammonium hydroxide. The light sensitizer $LaFeO_3$ was prepared by employing a sol-gel method. The flow chart for stepwise preparation of perovskite material is shown below in Fig. 2. The lanthanum and iron-based solution were prepared for synthesis of $LaFeO_3$. Deionized (DI) water chosen as a solvent to dissolve both

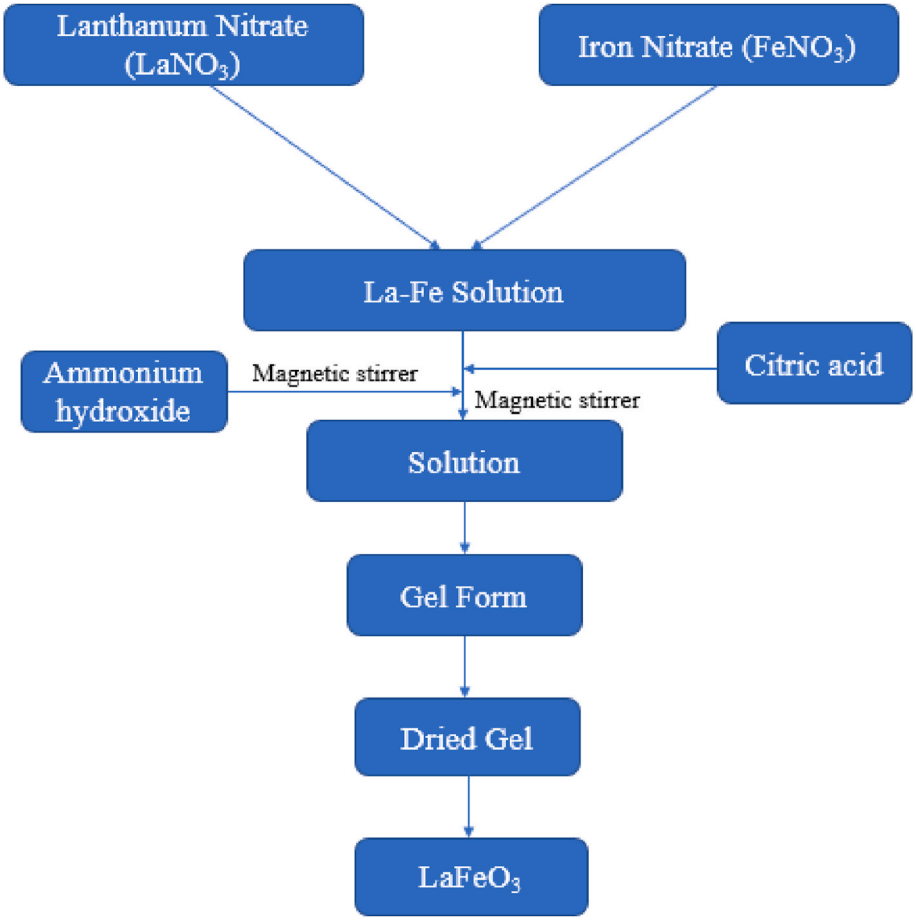


Fig. 2. Flow chart of stepwise synthesis of $LaFeO_3$.

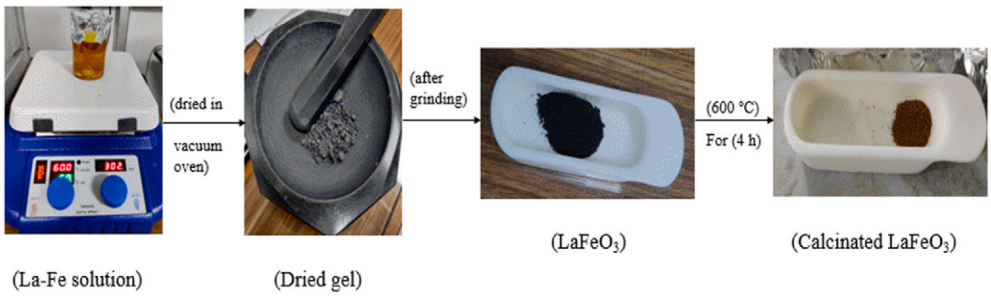


Fig. 3. Preparation of $LaFeO_3$ perovskite.

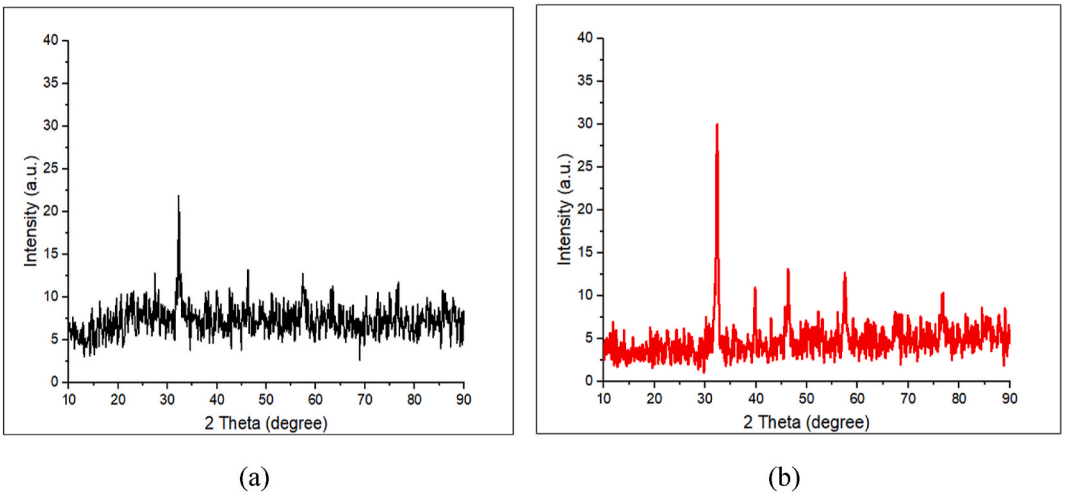


Fig. 4. XRD pattern of (a) $LaFeO_3$ and (b) Calcined $LaFeO_3$.

solutes. At first,1 M aqueous solution of $La(NO_3)_3$ mixed with 1 M aqueous solution of $Fe(NO_3)_3$. The mixture was stirred for 2 days at 300 rpm and 60 °C. Citric acid was added dropwise in the solution. After that, ammonium hydroxide aqueous solution was added dropwise under magnetic stirring to make pH = 7. The sample was stirred at 300 rpm for 72 h and dried in a vacuum oven at 180 °C for 12 h. After that sample was moved into a muffle furnace for calcination in air at 600 °C for 4 h. The synthesized sample shown in Fig. 3.

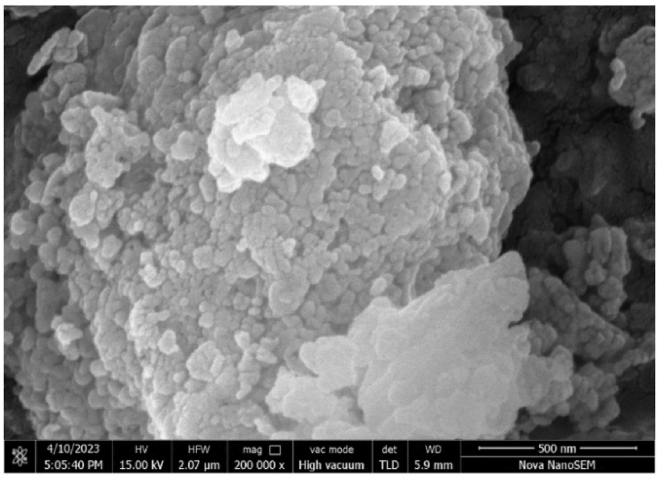


Fig. 5. SEM images of $LaFeO_3$ nanoparticles.

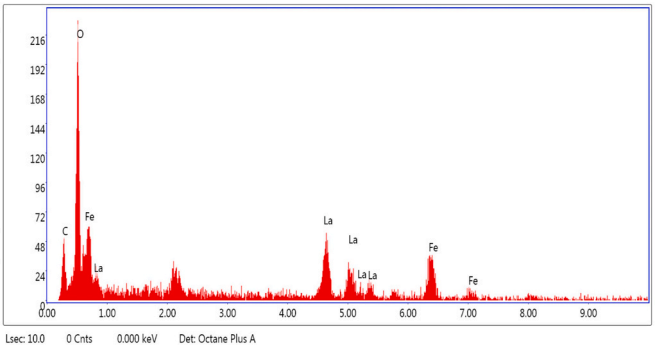


Fig. 6. EDS spectrum of $LaFeO_3$ nanoparticles.

3.2. Characterizations

XRD pattern of perovskite material was recognized utilizing a diffractometer (Bruker D8 advance). The size and morphologies of synthesized perovskite by SEM. UV–vis spectrum of the $LaFeO_3$ sample is observed by employing a PerkinElmer, Lamda-750 dual-beam spectrometer. The prepared sample IR transmission spectra was analyzed by employing an FTI-R spectrometer.

Table 1
Weight % of the element composition of $LaFeO_3$ powder.

Element	Series	Weight %	Atomic %
C	K-Series	6.70	22.42
O	K-Series	17.18	43.16
La	L-Series	47.37	13.71
Fe	K-Series	28.75	20.70

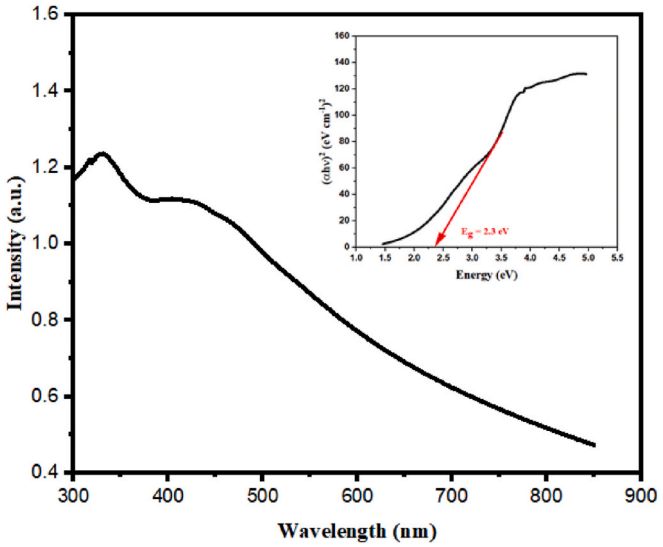


Fig. 7. UV–Vis spectrum of $LaFeO_3$ nanoparticles.

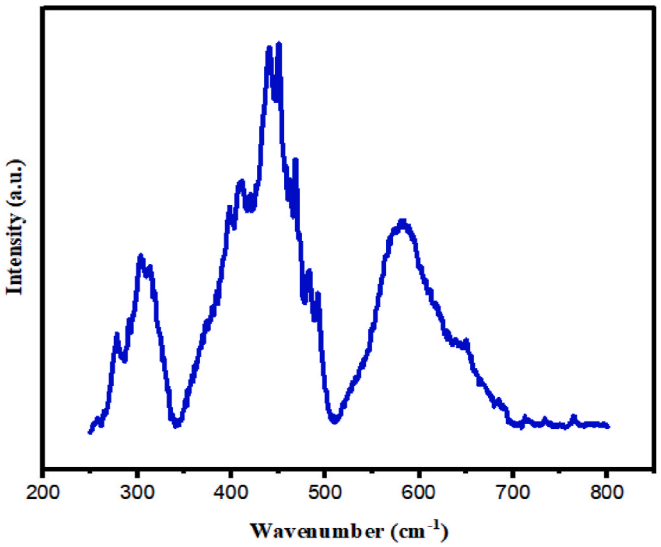


Fig. 9. PL spectrum of $LaFeO_3$ nanoparticles.

4. Experimental results and discussion

4.1. XRD analysis

It is primarily used for phase identification and crystallinity of the material. The observed range of XRD data is from 10° to 90° over 2θ with scan rate of 2°/min. The peaks of $LaFeO_3$ before and after calcination are demonstrated in Fig. 4 (a) and (b), respectively. The XRD peaks with 2θ of 32.30°, 39.91°, 46.42°, 57.60°, 67.83° and 76.68° correspond to $LaFeO_3$ perovskite nanoparticles of cubic structure (75–0541). Bragg's equation (1) was used to calculate the material grain size and display angles.

$$2d\sin\theta = n\lambda \quad (1)$$

$$\text{Crystallite Size} = (0.94 \times \lambda)/\ \beta\cos\theta \quad (2)$$

Where, λ = wavelength of X-ray, β = FWHM, θ = Bragg's angle.

The calculated average crystallite size is 18.4 nm by using Debye Scherrer's Eq. (2). It is observed that the peak intensities increase after the calcination (600 °C for 4 h) of the $LaFeO_3$, which confirms the uniform crystal growth of perovskite phase [21].

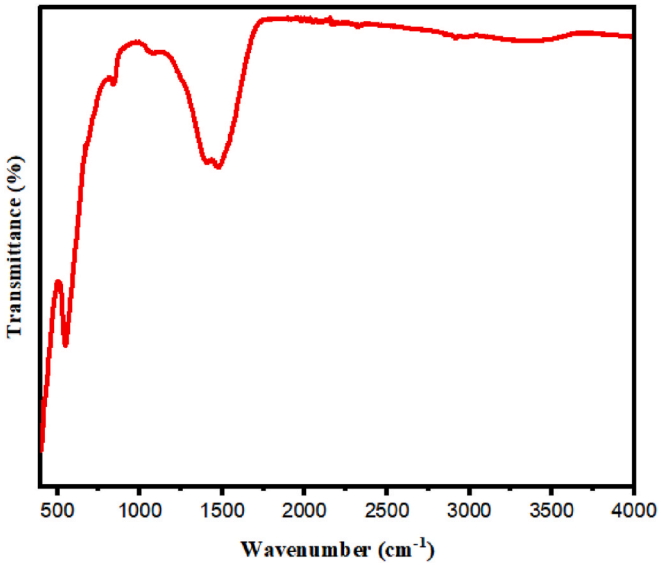


Fig. 8. FT-IR spectrum of $LaFeO_3$ nanoparticles.

Table 2
Input parameters of PSCs.

Parameter	FTO [29, 30]	TiO_2 [31, 32]	$LaFeO_3$ [20, 33]	Cu_2O [29, 33]
Thickness (nm)	500	100	400	50
Eg (eV)	3.5	3.2	2.3	2.17
χ (eV)	4.0	3.9	3.3	3.2
ε	9.0	32	30	7.11
N_C (cm^{-3})	2.2×10^{18}	1×10^{19}	4×10^{18}	2.02×10^{17}
N_V (cm^{-3})	1.8×10^{19}	1×10^{19}	1×10^{19}	1.1×10^{19}
μ_n (cm^2/Vs)	20	20	1.1×10^{-1}	2×10^{2}
μ_p (cm^2/Vs)	10	10	1.1×10^{-3}	80
N_D (cm^{-3})	1×10^{15}	1×10^{17}	1×10^{15}	0
N_A (cm^{-3})	0	0	1×10^{15}	1×10^{18}
N_t (cm^{-3})	1×10^{15}	1×10^{13}	1×10^{13}[a]	1×10^{14}

[a] In this work.

Table 3
Input parameters of different ETLs.

Parameter	PCBM [35]	IGZO [20, 36]	In_2O_3 [20]	C_{60} [30]	ZnO [37]	SnO_2 [31, 37]	WS_2 [38]
Thickness (μm)	100[a]	100[a]	100[a]	100[a]	100[a]	100	100
Eg (eV)	2.1	3.05	2.8	1.7	3.3	3.3	1.8
χ (eV)	3.9	4.16	3.7	3.9	4.1	4.0	3.95
ε	3.9	10	8.9	4.2	9.0	9.0	13.6
N_C (cm^{-3})	2.2×10^{19}	5×10^{18}	1×10^{19}	8×10^{19}	2.2×10^{18}	2.2×10^{17}	1×10^{18}
N_V (cm^{-3})	2.2×10^{19}	5×10^{18}	1×10^{19}	8×10^{19}	1.9×10^{19}	2.2×10^{16}	2.4×10^{19}
μ_n (cm^2/Vs)	1×10^{-3}	15	3.4	8×10^{-2}	100	200	100
μ_p (cm^2/Vs)	2×10^{-3}	0.1	3.4	3.5×10^{-3}	25	80	100
N_D (cm^{-3})	1×10^{18}	1×10^{18}	1×10^{18}	2.6×10^{17}	1×10^{17}	1×10^{17}	1×10^{18}
N_A (cm^{-3})	0	0	0	0	0	0	0
N_t (cm^{-3})	1×10^{13}[a]	1×10^{13}[a]	1×10^{13}[a]	1×10^{13}[a]	1×10^{13}[a]	1×10^{13}[a]	1×10^{13}[a]

[a] In this work.

Table 4
Input parameters of different HTLs.

Parameter	$CuInSe_2$ [39,40]	Spiro-OMeTAD [29,33,39,41]	P3HT [30,42]	D-PBTTT-14 [29]	V_2O_5 [39,43]	CFTS [44]	CBTS [44]
Thickness (μm)	50[a]	50[a]	50	50	50[a]	50[a]	50[a]
Eg (eV)	2.17	3.0	1.7	2.16	2.2	1.3	1.9
χ (eV)	3.2	2.45	3.5	3.2	4.0	3.3	3.6
ε	6.6	3.0	3.0	10	10	9.0	5.4
N_C (cm^{-3})	2.5×10^{20}	1×10^{19}	2×10^{18}	2.8×10^{19}	9.2×10^{17}	2.2×10^{18}	2.2×10^{18}
N_V (cm^{-3})	2.5×10^{20}	1×10^{19}	2×10^{19}	1×10^{19}	5×10^{18}	1.8×10^{19}	1.8×10^{19}
μ_n (cm^2/Vs)	80	1×10^{-4}	1.8×10^{-3}	2.83×10^{-3}	3.2×10^{2}	21.98	30
μ_p (cm^2/Vs)	80	1×10^{-4}	1.8×10^{-2}	2.83×10^{-3}	40	21.98	10
N_D (cm^{-3})	0	0	0	0	0	0	0
N_A (cm^{-3})	1×10^{18}	2×10^{18}	1×10^{18}	1×10^{18}	1×10^{18}	1×10^{18}	1×10^{18}
N_t (cm^{-3})	1×10^{14}[a]	1×10^{14}	1×10^{14}	1×10^{14}	1×10^{14}	1×10^{14}[a]	1×10^{14}[a]

[a] In this work.

Table 5
The PSC performance of CBTS with the different ETLs.

Device architecture	V_{OC} (V)	J_{SC} (mA/cm^2)	FF (%)	PCE (%)
FTO/TiO_2/$LaFeO_3$/CBTS/Au	1.82	6.32	68.13	7.87
FTO/PCBM/$LaFeO_3$/CBTS/Au	1.76	7.53	77.49	10.32
FTO/IGZO/$LaFeO_3$/CBTS/Au	1.87	6.41	57.19	6.88
FTO/In_2O_3/$LaFeO_3$/CBTS/Au	1.82	6.37	81.49	9.46
FTO/C_{60}/$LaFeO_3$/CBTS/Au	1.37	11.36	48.59	7.62
FTO/ZnO/$LaFeO_3$/CBTS/Au	1.82	6.38	61.62	7.17
FTO/SnO_2/$LaFeO_3$/CBTS/Au	1.81	6.29	69.20	7.92
FTO/WS_2/$LaFeO_3$/CBTS/Au	1.59	10.34	76.11	12.53

4.2. SEM results

The SEM is employed to inspect topography and superficial characteristics of $LaFeO_3$ perovskite material. The nanoparticles of $LaFeO_3$ did not aggregate and have small size depicted in Fig. 5. The nanoparticles are discrete with an average size of ~80 nm. The SEM investigation showed that film is not uniform due to there are no surface-protecting layers which lead to agglomeration of the particles. The particle shapes are ambiguous, numerous large and small pores exist in the entire material [15].

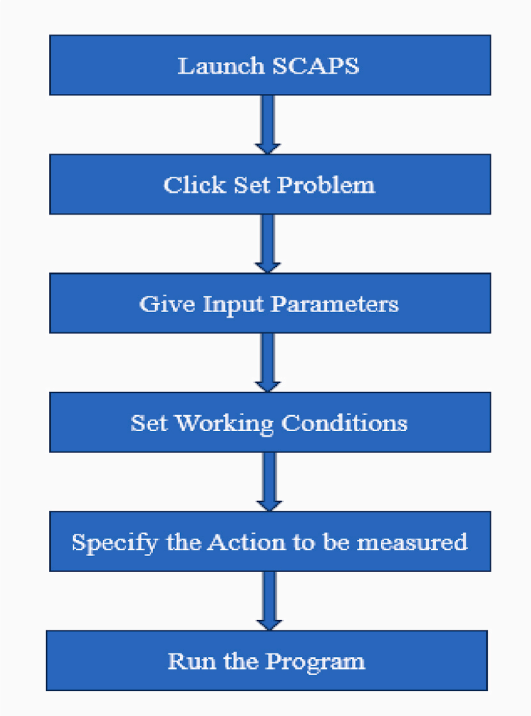


Fig. 10. Simulation procedure.

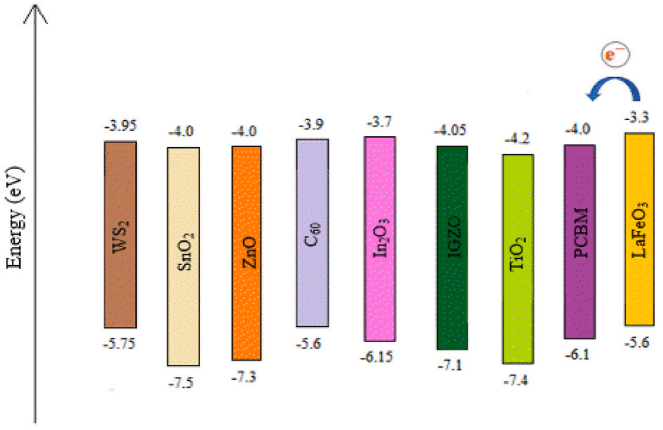


Fig. 11. Band alignment diagram of various ETLs with $LaFeO_3$ absorber layer.

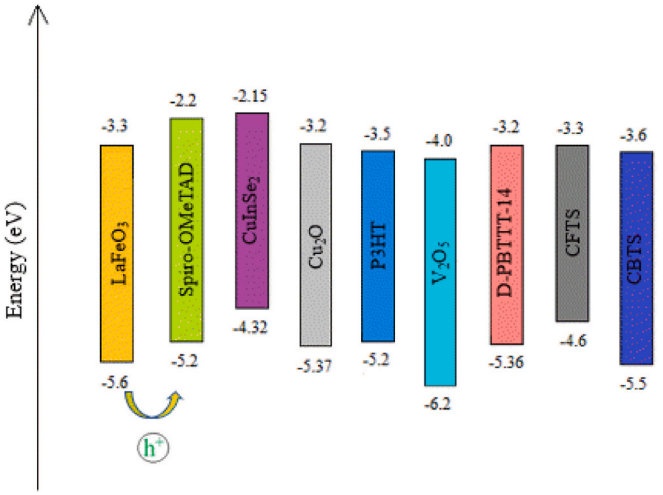
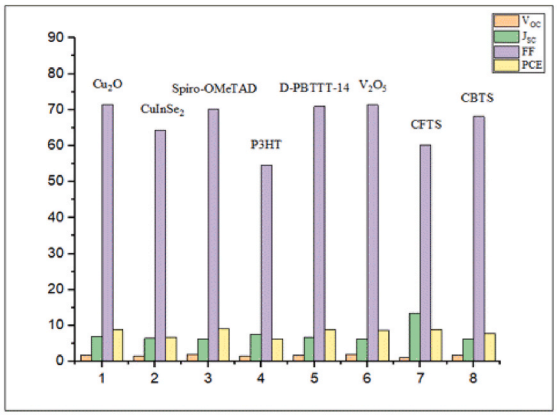


Fig. 12. Band alignment diagram of various HTLs with $LaFeO_3$ absorber layer.

4.3. EDS spectrum

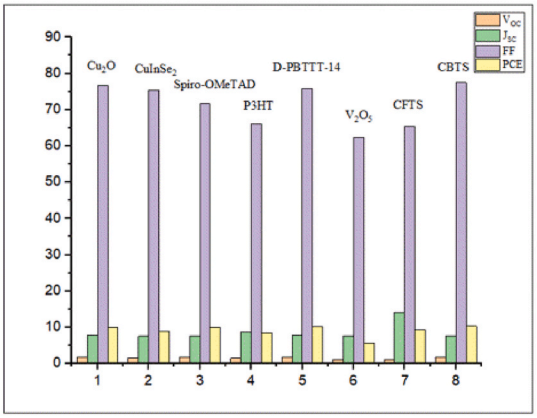
The composition and chemical purity of elements are investigated using EDS. Fig. 6 demonstrates EDS spectroscopy of $LaFeO_3$. The diagram shows that La, Fe, C and O are observed without no peaks of impurities. The presence of these elements can validate the formation of pure phases. The weight percent of all elements presented in $LaFeO_3$ perovskite were shown below in Table 1.

(a) TiO_2



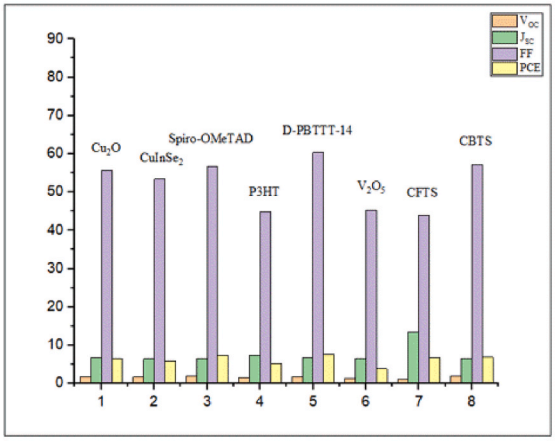
Voc	1.77	1.63	2.00	1.52	1.84	1.92	1.09	1.82
Jsc	6.94	6.42	6.38	7.60	6.72	6.38	13.47	6.32
FF	71.47	64.30	70.24	54.62	70.92	71.35	60.24	68.13
PCE	8.81	6.75	9.00	6.34	8.82	8.77	8.92	7.87

(b) PCBM



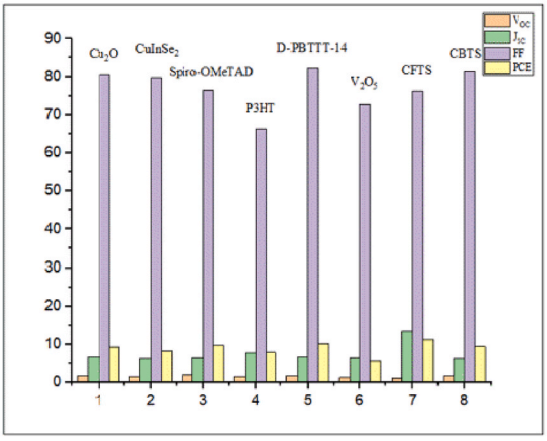
Voc	1.66	1.58	1.86	1.47	1.74	1.20	1.00	1.76
Jsc	7.72	7.53	7.56	8.67	7.74	7.54	14.10	7.53
FF	76.71	75.32	71.60	66.11	75.81	62.33	65.50	77.49
PCE	9.89	8.98	10.07	8.47	10.23	5.65	9.33	10.32

(c) IGZO



Voc	1.77	1.69	2.01	1.57	1.86	1.30	1.11	1.87
Jsc	6.70	6.41	6.47	7.48	6.74	6.46	13.54	6.41
FF	55.74	53.41	56.66	44.91	60.31	45.32	43.95	57.19
PCE	6.63	5.82	7.40	5.29	7.58	3.81	6.67	6.88

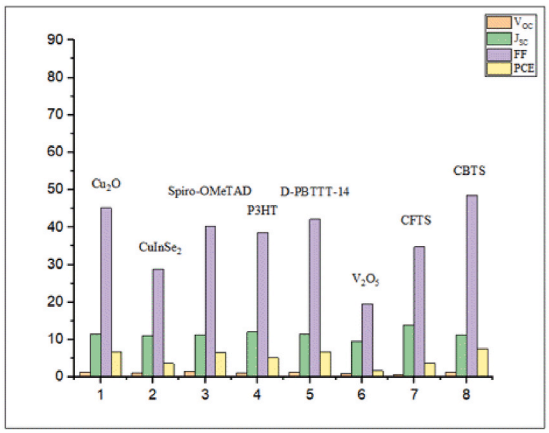
(d) In_2O_3



Voc	1.72	1.62	2.00	1.52	1.83	1.22	1.09	1.82
Jsc	6.57	6.39	6.41	7.86	6.68	6.41	13.49	6.37
FF	80.63	79.72	76.51	66.21	82.34	72.83	76.24	81.49
PCE	9.28	8.30	9.84	7.93	10.12	5.73	11.30	9.46

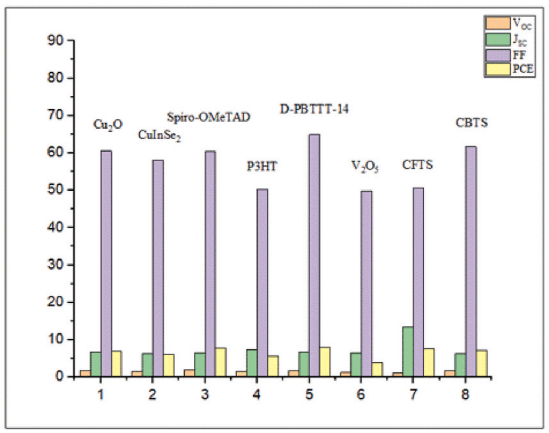
Fig. 13. Optimization of $LaFeO_3$ PSC PV parameters for various ETLs (TiO_2, PCBM, IGZO, In_2O_3, C_{60}, ZnO, SnO_2, WS_2) and HTLs (Cu_2O, $CuInSe_2$, Spiro-OMeTAD, P3HT, D-PBTTT-14, V_2O_5, CFTS, CBTS) with Au as metal back contact.

(e) C_{60}



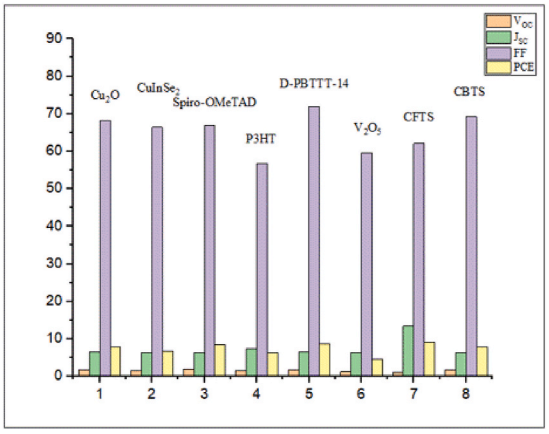
Voc	1.29	1.13	1.44	1.10	1.37	0.89	0.67	1.37
Jsc	11.49	11.10	11.36	12.03	11.50	9.58	13.93	11.36
FF	45.21	28.84	40.29	38.62	42.18	19.61	34.82	48.59
PCE	6.72	3.64	6.61	5.13	6.69	1.68	3.26	7.62

(f) ZnO



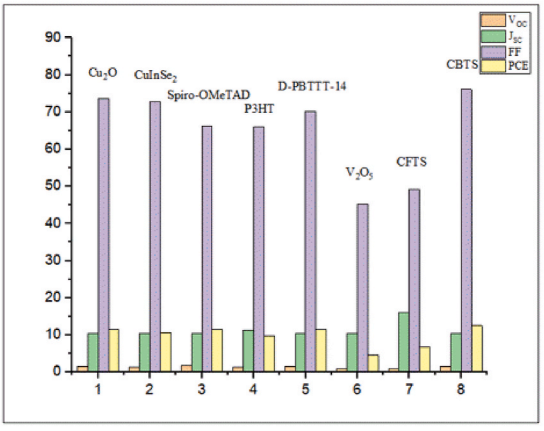
Voc	1.72	1.62	2.00	1.52	1.83	1.22	1.09	1.82
Jsc	6.67	6.38	6.45	7.34	6.71	6.43	13.53	6.38
FF	60.58	58.06	60.44	50.18	64.88	49.82	50.71	61.61
PCE	6.97	6.04	7.81	5.62	8.01	3.93	7.53	7.17

(g) SnO_2



Voc	1.72	1.62	2.00	1.52	1.83	1.22	1.09	1.81
Jsc	6.58	6.29	6.35	7.42	6.61	6.34	13.44	6.29
FF	68.30	66.38	66.97	56.67	71.98	59.60	62.02	69.20
PCE	7.75	6.80	8.52	6.40	8.75	4.63	9.15	7.92

(h) WS_2



Voc	1.49	1.41	1.66	1.30	1.56	0.99	0.84	1.59
Jsc	10.50	10.35	10.37	11.29	10.51	10.38	16.12	10.34
FF	73.65	72.70	66.22	66.03	70.18	45.21	49.16	76.11
PCE	11.59	10.62	11.42	9.74	11.58	4.66	6.69	12.53

Fig. 13. (*continued*).

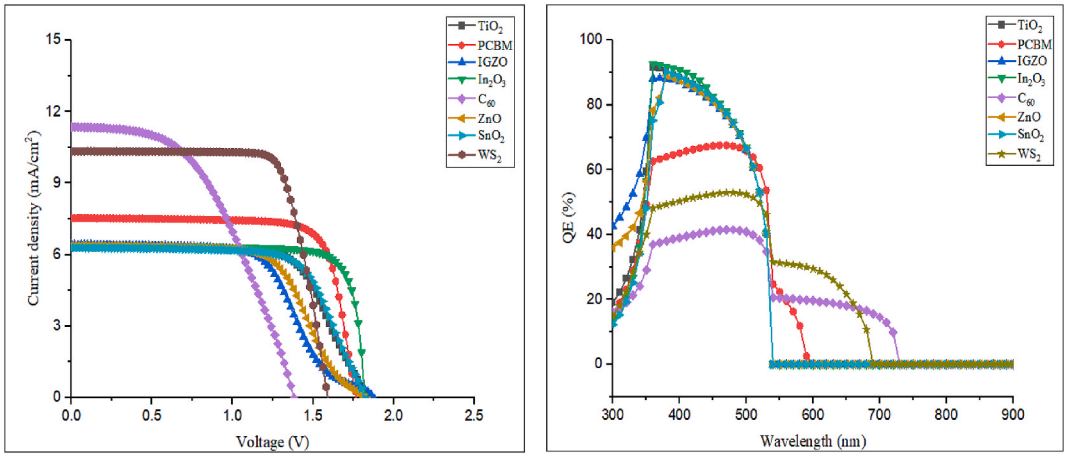


Fig. 14. J-V and QE with distinct ETLs.

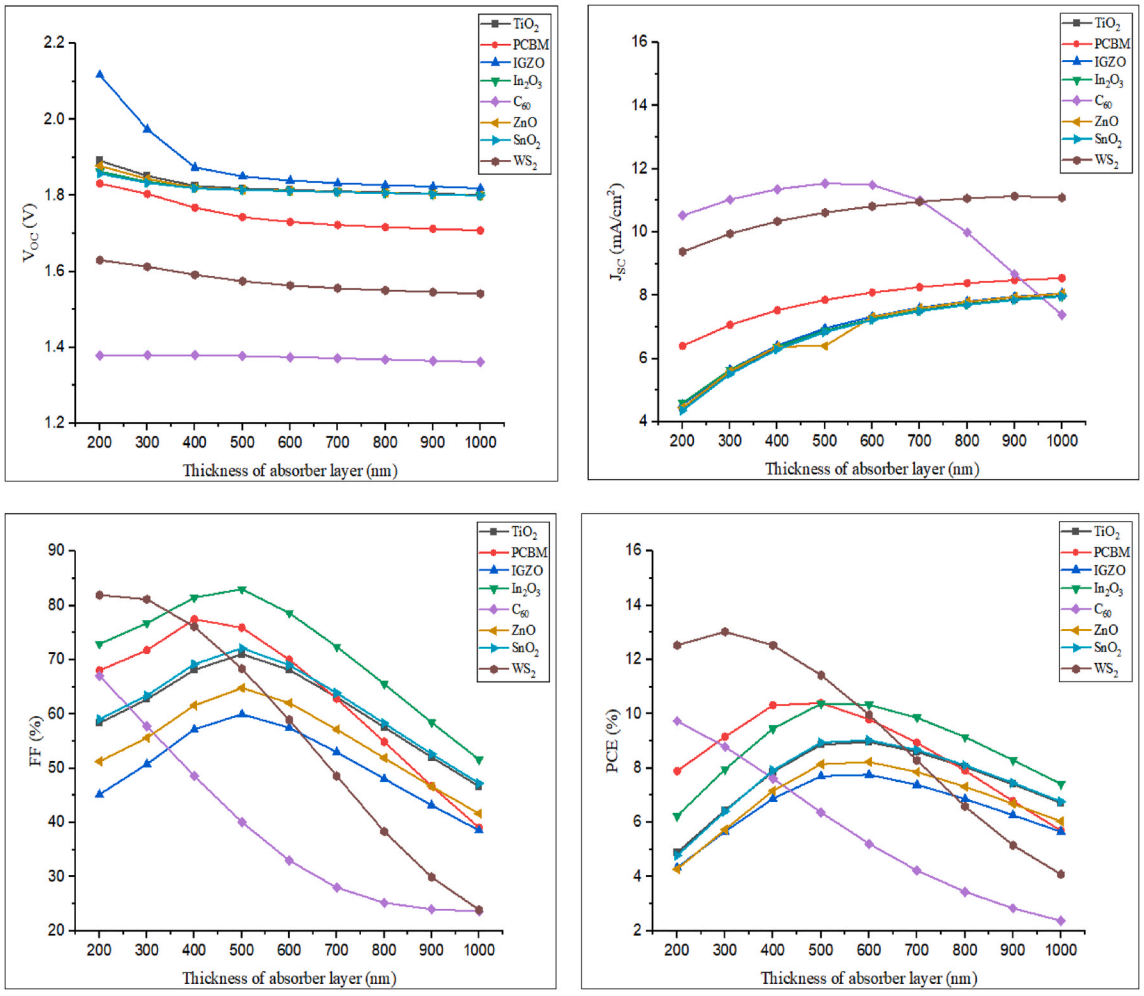


Fig. 15. Impact of thickness on PV parameters of $LaFeO_3$-based PSCs.

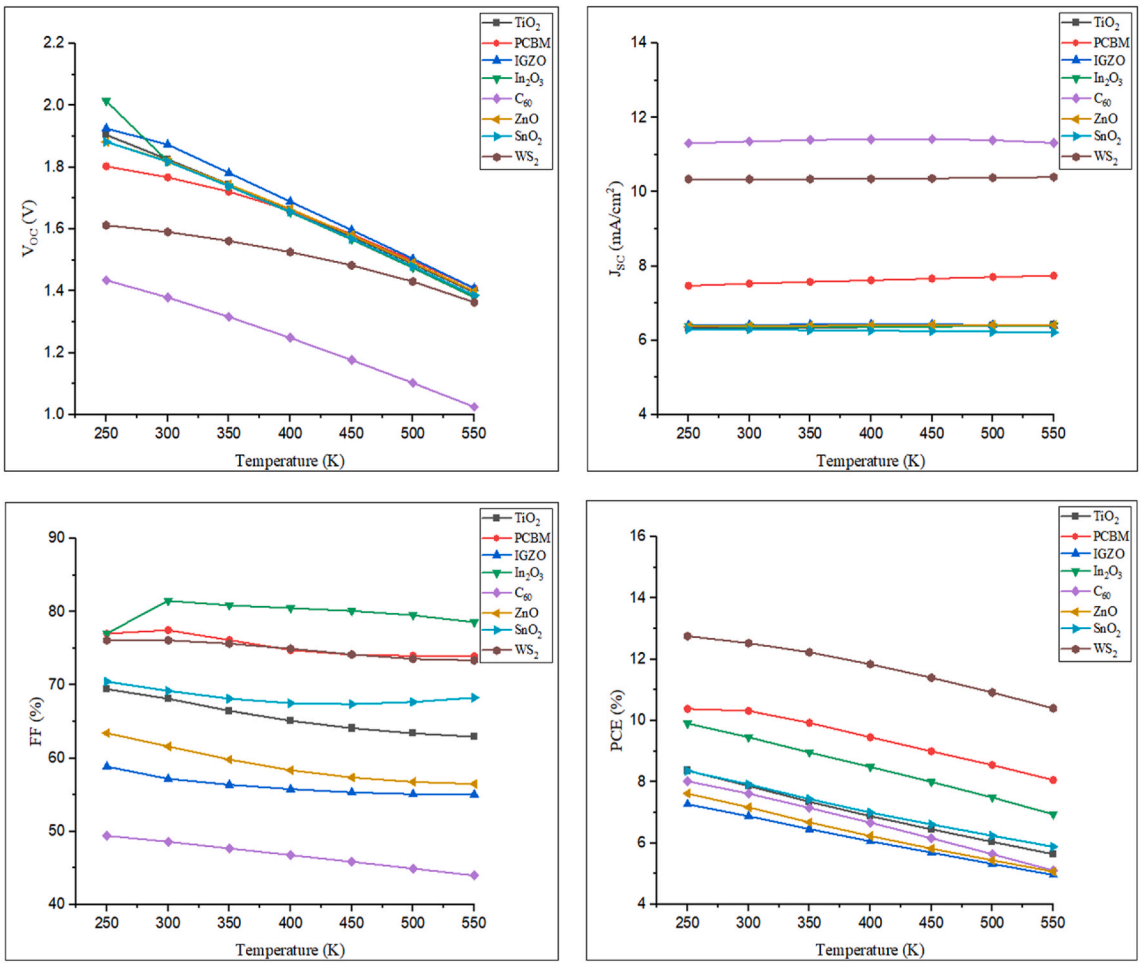


Fig. 16. Change in performance of $LaFeO_3$-based PSCs with operating temperature.

4.4. UV–vis spectroscopy

UV–Vis spectrum of $LaFeO_3$ perovskite material carried out in order to calculate optical E_g illustrated in Fig. 7. The highest absorption peak of the $LaFeO_3$ sample is observed at 330 nm. The E_g is calculated by equation (3).

$$E_g = \frac{hc}{\lambda_{max}} \tag{3}$$

Where, E_g is the band gap energy of the sample, h = Planck's constant, c = speed of light. λ_{max} is the extreme absorption wavelength of the sample. The calculated value of E_g for $LaFeO_3$ perovskite is approximately 2.3 eV and is in range as reported by Azozi et al. [23].

4.5. FT-IR analysis

FT-IR is employed to distinguish the vibrational modes of $LaFeO_3$ perovskite material and confirmation of existent functional groups. The FT-IR spectrum of a $LaFeO_3$ perovskite in the range of 400–4000 cm^{-1} presented in Fig. 8. The peaks of the $LaFeO_3$ nanoparticles are near to 552, 838, 1061, 1404, 1480, 2912 and 3359 cm^{-1}. The absorption band is observed at 552 cm^{-1} which is attributed to Fe–O stretching vibration mode [24,25]. This peak confirms the formation of perovskite phase which commonly exists between 500 and 700 cm^{-1} [26]. The IR broad peak at 3359 cm^{-1} arises from H–O bending mode and 2109 cm^{-1} due to symmetric stretching of the carboxyl group [27].

4.6. Photoluminescence spectrum

To delve deeper into the electronic properties, we conducted room temperature Photoluminescence spectrum (PL) analysis on the synthesized powder. In this investigation, $LaFeO_3$ exhibited a prominent emission peaking around 450 nm, aligning with the band-edge emission observed with excitation wavelength of 330 nm shown below in Fig. 9. The visible emission bands might be associated with oxygen vacancies serving as luminescent sites. These vacancies naturally occur in oxygen-deficient environments and manifest as defects on the surface or within the lattice structures. In this scenario, these oxygen vacancies serve as luminescent centers, capturing both a h^+ and an e^- on the nanostructures surface. The recombination of this e^-- h^+ pair results in nonradiative emission. The oxygen vacancies found in synthesized perovskites prompt further investigation into their electrical and optical characteristics.

5. Simulation methodology

5.1. SCAPS-1D

The data of different parameters are chosen from previous studies. The simulation was carried out using SCAPS-1D (version 3.3.10). Gent University of Belgium developed SCAPS-1D software, enabling users to model up to seven semiconducting layers under both light and dark conditions. This software can be used for simulations at different input physical conditions. The tool calculation is based on Poisson's and continuity equations for both e^-s and h^+s under steady-state conditions

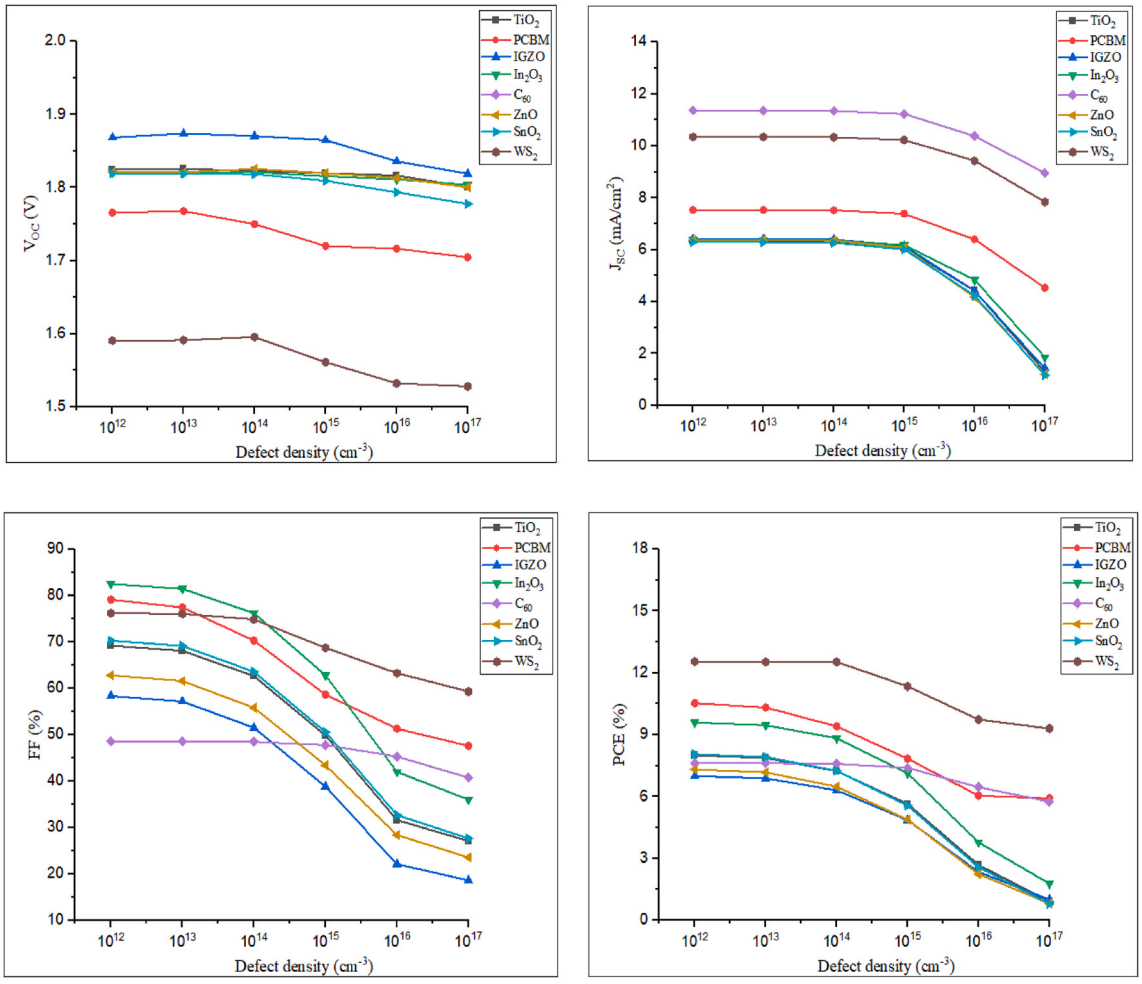


Fig. 17. Impact of N_t on the performance of $LaFeO_3$-based PSCs.

Table 6
Defect properties of $LaFeO_3$ absorber layer.

Parameters	Values
Density of defect, N_t (cm^{-3})	10^{12}–10^{17}
Capture cross section electron σ_n (cm^2)	1×10^{-15}
Capture cross section hole σ_h (cm^2)	1×10^{-15}
Energetic distribution	Single
Type	Neutral
Energy level	Midgap

as described below [28].

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{q}{\epsilon}[p(x) - n(x) + N_D - N_A + \rho_p - \rho_n = 0 \quad (4)$$

$$\frac{1}{q}\frac{dJ_p}{dx} = G_{op}(x) - R(x) \quad (5)$$

$$\frac{1}{q}\frac{dJ_n}{dx} = -G_{op}(x) + R(x) \quad (6)$$

The PSC consists of $LaFeO_3$ perovskite absorber material with layer architecture: glass substrate/FTO/ETL/$LaFeO_3$/HTL/Back contact depicted in Fig. 1. The input parameters employed in simulation of several layers are enclosed in Tables 2–4. The simulation procedure is shown in Fig. 10Table 5.

6. Simulation results and discussion

6.1. Optimization of suitable ETL and HTL

The various ETLs such as TiO_2, PCBM, IGZO, In_2O_3, C_{60}, ZnO, SnO_2 and WS_2 were inserted into the heterostructure to achieve the optimum performance. The band alignment of various ETLs demonstrated in Fig. 11. The input parameters of employed ETLs are demonstrated in Table 3. We have done the simulation by varying the thickness of all ETLs from 50 to 250 nm. We found that the optimized value of ETL thickness is 100 nm for all ETLs. It has been analyzed that there is a negligible change in efficiency of the $LaFeO_3$-based PSC for values greater or less than the optimized value. The different configurations of eight ETLs and eight HTLs are studied to achieve the maximum performance. The band matching is a vital factor. One of the critical issues confronting the PSCs community involves the alignment of energy level between the ETLs and the absorber layer. To tackle this issue, a thorough examination of the ETL/perovskite interface is imperative. Defects arising from lattice mismatching at the ETL/perovskite interface play a vital role in trapping the charge carriers and greatly influence the device efficiency. The conduction band of the WS_2 layer is well aligned to the lowest unoccupied molecular orbital (LUMO) of the absorber layer. This can facilitate smooth charge carrier transport, minimizing the impact of defects that may otherwise trap or recombine charges. This band alignment ensures that e^-s can easily move from the absorber layer to the WS_2 layer and eventually to the contact, contributing to higher overall device efficiency [34]. The conduction band of the ETL is always less than the LUMO of the absorber layer. The gap amid the valence band of

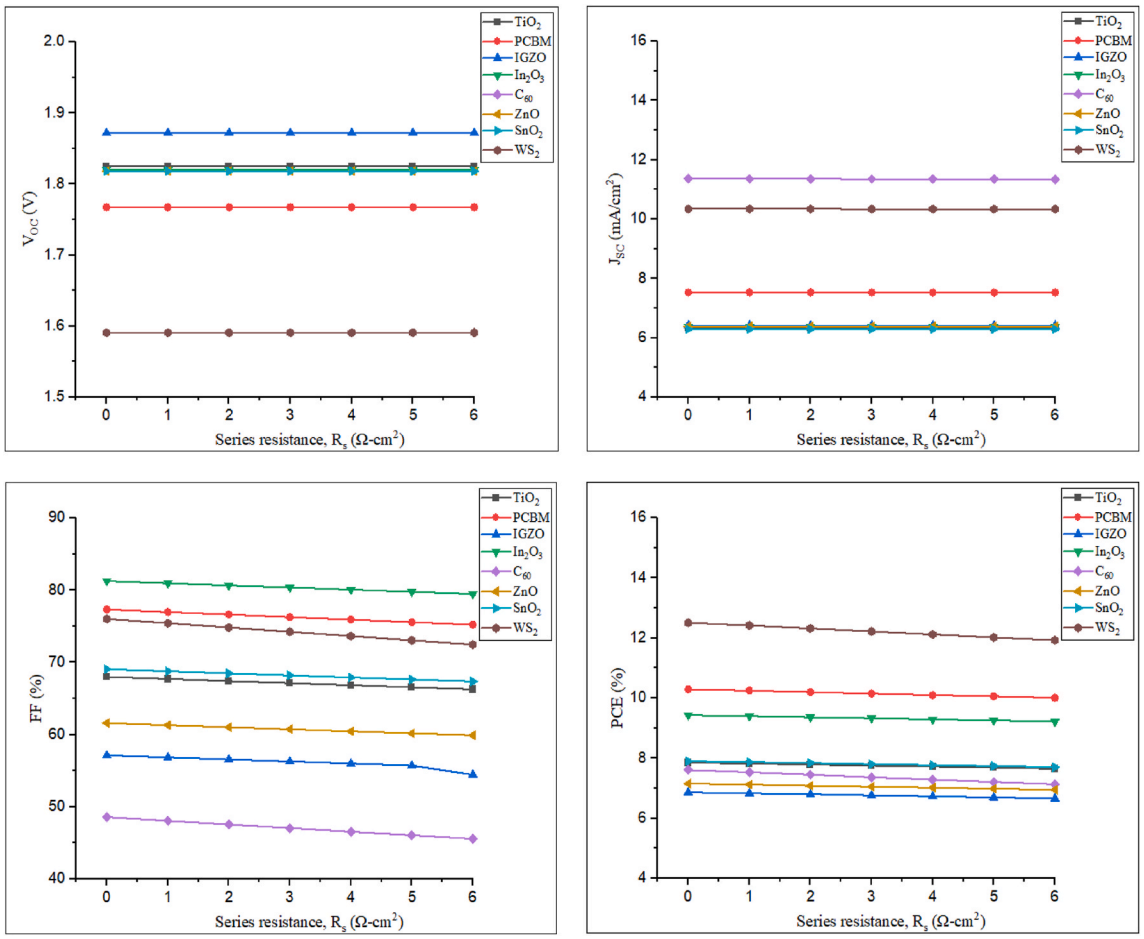


Fig. 18. Impact of R_S on the performance of $LaFeO_3$-based PSC at $R_{SH} = 10^5$ Ω-cm^2.

ETL and HOMO level of the absorber layer must always maintain substantial distance, which inhibits the recombination in the absorber layer. The results illustrate that TiO_2, SnO_2 and ZnO have wide E_g, so that they have exhibited low PCE of around 8 %. ETLs like IGZO, and C_{60} suffering from mismatch in band alignment hence, exhibited low efficiency relatively. In this work WS_2 is most suitable ETL to achieve maximum PCE for $LaFeO_3$-based PSC. The WS_2 has advantages such as band matching, photoelectrochromic, toughness, high melting point and corrosion resistance. Due to these properties, WS_2 is the promising candidate for optoelectronic application. The maximum efficiency of 13.03 % has been obtained with the heterostructure of FTO/WS_2/$LaFeO_3$/CBTS/Au.

Impact of distinct HTLs such as Cu_2O, $CuInSe_2$, Spiro-OMeTAD, P3HT, D-PBTTT-14, V_2O_5, CFTS and CBTS with eight different combinations were studied. In contrast, the V_2O_5, and CFTS achieved less efficiency than other HTLs. The remarkable enhancement in efficiency of a PSC with the inorganic HTL. The inorganic HTL materials are more stable, transparent, and band-aligned, over organic HTL. CBTS is most suitable HTL for $LaFeO_3$-based PSC due to its diverse properties such as crystalline structure, light absorption, atomic size and an earth-abundant material [22]. The CBTS as HTL offered most suitable band matching as compared to other HTLs shown in Fig. 12, so that it exhibited the highest PCE of 13.03 % for $LaFeO_3$-based PSCs. The total sixty-four different heterostructures have been simulated to explore appropriate combinations of ETL and HTL for $LaFeO_3$-based PSC to achieve maximum performance depicted in Fig. 13. The J-V and QE plots of the best eight configurations are illustrated in Fig. 14.

6.2. Impact of thickness of absorber layer

The thickness altered from 200 to 1000 nm at N_t of 1×10^{13} cm^{-3} and temperature of 300 K. It is known to us that as thickness of the layer decreases, absorption rate also lowers. This results in a decrease in photocurrent and therefore PCE of the solar cell also reduces after 300 nm. As an increase in thickness, large numbers of photons are absorbed but simultaneously recombination rate increases and high-rate results in V_{OC} [29,33]. Both V_{OC} and PCE parameters decrease exponentially for increase in thickness from 300 to 1000 nm. It can be seen that WS_2 PCE is highest at around 300 nm. Whereas for C_{60} there is a rapid fall in PCE due to very low value of electron mobility as compared to other ETLs. Hence it can be deduced that the Optimum value of thickness is 300 nm for $LaFeO_3$-based PSC and beyond that PCE falls and decrease in FF is 68.61 %. The maximum efficiency of PSC goes up to 13.03 % together with V_{OC} is 1.61 V, J_{SC} is 9.952 mA/cm^2 and FF is 81.17 % for the thickness of 300 nm at ETL = 100 nm and HTL = 50 nm. The FF shows a monotonic decrement with the rise in thickness when thickness becomes greater than 300 nm. The change in different PV parameters with thickness of $LaFeO_3$ layer with CBTS as HTL depicted in Fig. 15.

6.3. Effect of operating temperature

The working temperature range is between 250 and 550 K. Fall in the performance with temperature increase is related to a reduction in diffusion length of charge carriers. Several solar cell architectures show the instability of their overall performance due to deformation among layers at high temperatures. In $LaFeO_3$-based PSC, fall in efficiency with temperature is 18.11 % for FTO/WS_2/$LaFeO_3$/CBTS/Au configuration.

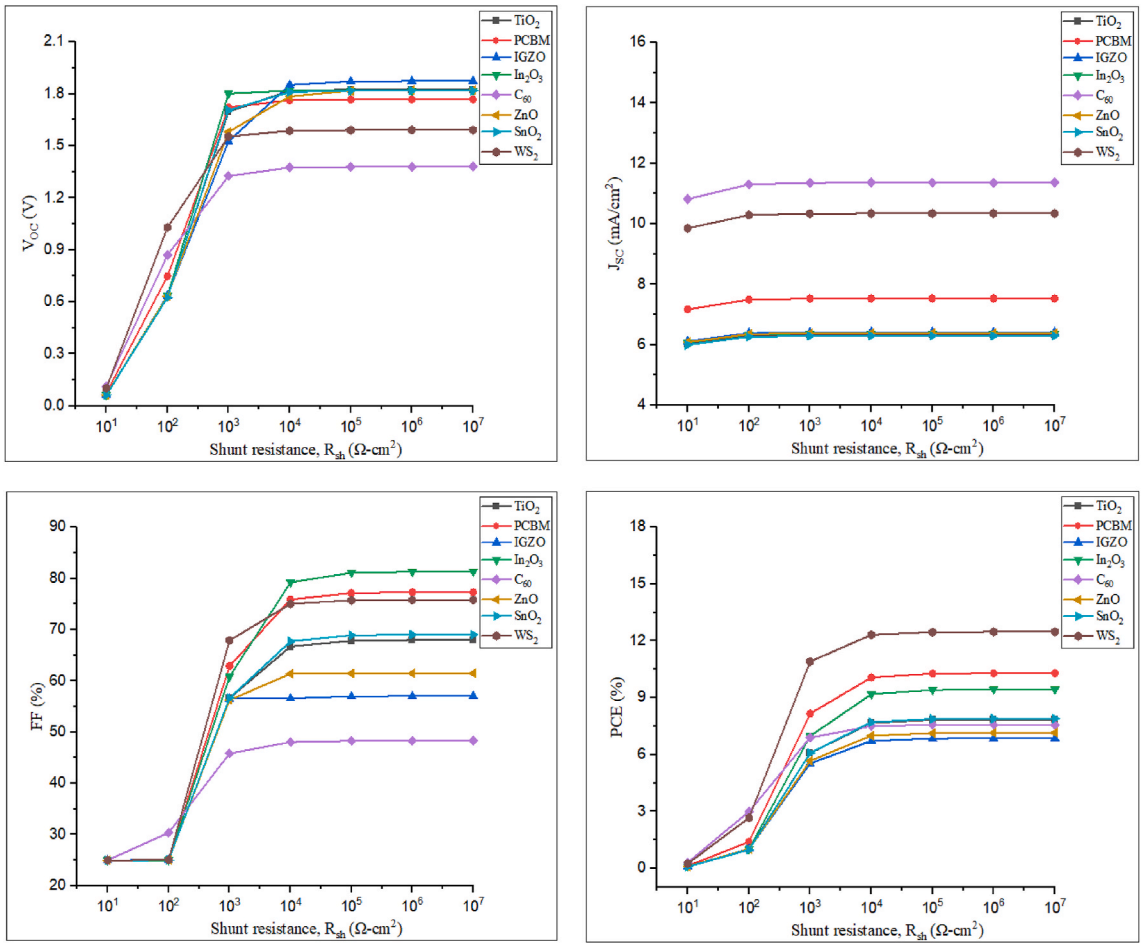


Fig. 19. Impact of R_{sh} on the performance of $LaFeO_3$-based PSC at $R_S = 0.5$ Ω-cm^2.

When the enhancement of temperature occurs, the PCE drops significantly from 12.70 % to 10.40 % for the given range of temperature. The optimized temperature is 250 K for $LaFeO_3$-based PSC. The value of V_{OC} is decreasing with rise in temperature for all ETLs. The increment in temperature results in reduced the performance of PSC owing to the saturation current and therefore the recombination rate increases [45]. The impact of operating temperature with CBTS as HTL and various ETLs on the PV parameters are shown below in Fig. 16.

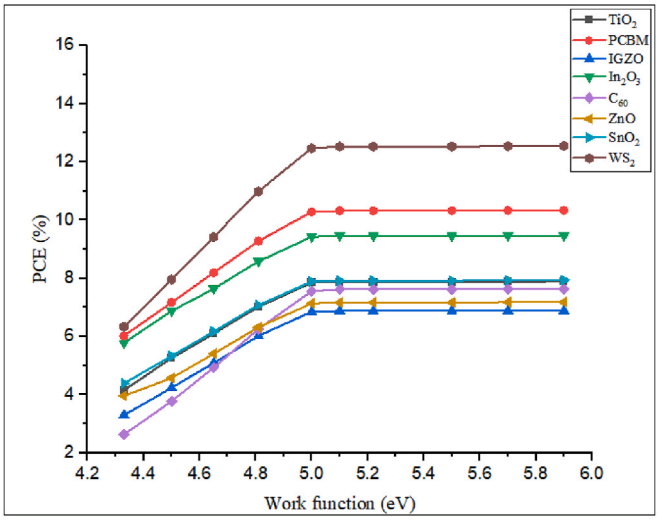


Fig. 20. Impact of various work functions on efficiency.

6.4. *Impact of defect density (N_t)*

The PSC performance and stability is expressively reduced at a large number of defects [46]. We have analyzed the PV parameters by varying N_t of the absorber layer from 10^{12} to 10^{17} cm^{-3}. The optimized value of N_t is 1×10^{13} cm^{-3} for $LaFeO_3$-based PSC at which it has achieved the maximum PCE of 12.53 %, as illustrated in Fig. 17. The PCE and J_{SC} fall sharply after 10^{15} cm^{-3} due to increase in defects hence number of traps of recombination enhances [47,48]. Increase in N_t, results in decrease in short circuit current, which is lastly responsible for fall in FF of the solar devices. The device PCE decreased from 12.53 % to 9.30 % for FTO/WS_2/$LaFeO_3$/CBTS/Au heterostructure, when N_t was enhanced from 10^{12} to 10^{17} cm^{-3}. Such an increasing amount of recombination resulted in a sharp decrease in PCE. The change in different PV parameters with N_t for $LaFeO_3$ with CBTS as HTL and different ETLs is depicted in Fig. 17. The fall in PCE is 25.89 % with an increase in value of N_t for optimized structure of $LaFeO_3$-based PSC. Some defect parameters of $LaFeO_3$ absorber layer shown below in Table 6.

6.5. *Effect of series resistance (R_s)*

The R_S has a significant impact on the PCE of the device specially on the FF and short circuit current (I_{SC}). With the increment in R_S, FF reduces which results in decrement in I_{SC}. Hence, devices exhibit poor performance due to large values of R_S. When solar cells are exposed to the environment, thermomechanical fatigue or cracks develop in the solder bonds depending on weather conditions. These cracks lead to an increase in the value of R_S which in turn causes a decline in the PCE [49,50]. The value of R_S varied from 0 to 6 Ω-cm^2, at R_{SH} of 10^5 Ω-cm^2. The

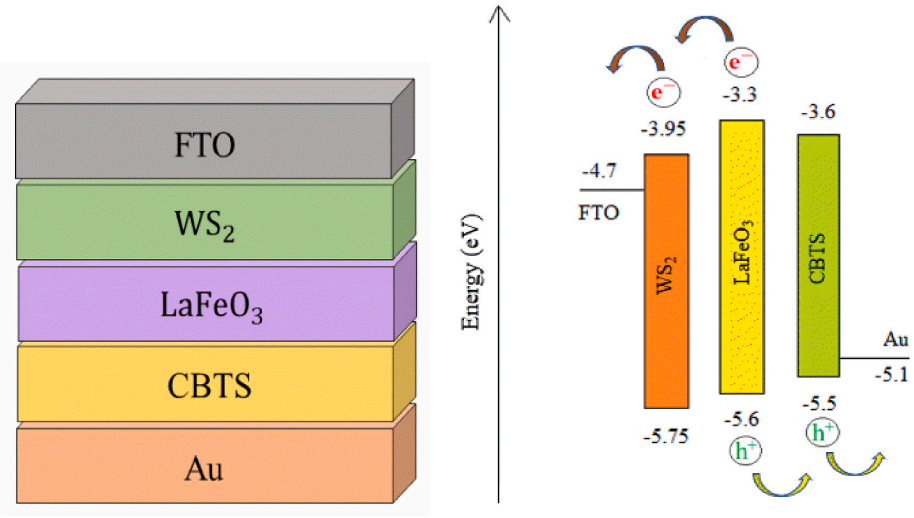


Fig. 21. Structure and band alignment diagram of optimized device.

Table 7
Comparison of Experimental, SCAPS-1D and previously reported outcomes.

Parameters	SCAPS-1D	Experimental	Experimental [21]	Simulation [21]	Simulation [20]
V_{OC} (V)	1.61	0.504	0.56	0.90	1.99
J_{SC} (mA/cm^2)	9.95	5.56	0.30	0.097	6.75
FF (%)	81.17	28.31	22	8.55	67.55
PCE (%)	13.03	0.79	0.06	0.01	9.11

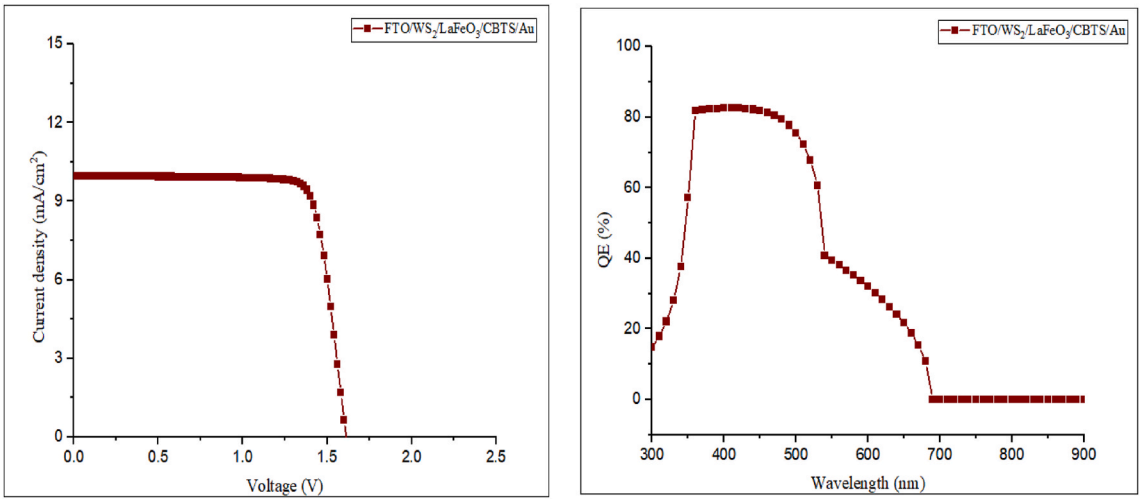


Fig. 22. J-V and QE curve of optimized device.

simulation results of change in value of R_S with PV parameters of PSC shown below in Fig. 18. The FF reduced from 76 to 72 % with an increase in value of R_S while the PCE, V_{OC} and J_{SC} were not affected in every device. The results shown in Fig. 18 are consistent with the previously reported studies [51,52]. The IGZO/$LaFeO_3$/CBTS and C_{60}/$LaFeO_3$/CBTS PSC configuration showed the highest V_{OC} and J_{SC} respectively whereas IGZO/$LaFeO_3$/CBTS PSC configuration showed relatively poor performance [53].

6.6. Effect of shunt resistance (R_{SH})

The R_{SH} varied from 10^1 to 10^7 Ω-cm^2 at constant value of R_S of 0.5 Ω-cm^2 for FTO/ETL/$LaFeO_3$/CBTS/Au configuration. The simulated outcome of different PV parameters shown below in Fig. 19. The J_{SC} varied from 10.82 to 11.36 mA/cm^2 for FTO/C_{60}/$LaFeO_3$/CBTS/Au structure whereas V_{OC} and PCE show a significant increment from 0.103 to 1.591 V and 0.25–12.48 % respectively for FTO/WS_2/$LaFeO_3$/CBTS/Au configuration. The small increase in J_{SC} with R_{SH} is due to decreased currents in the shunt-resisted path. The increment in V_{OC} is due to enhanced value of R_{SH} developing more voltage across the PSC [54]. The relation of the PCE with R_{SH} reported in the literature [55].

6.7. Appropriate back contact

The impact of back contact on PCE of a PSC is studied. We have simulated the different back contacts with a work function (Φ) of 4.3 eV–5.9 eV [56], the range covering the most of the available back contacts. The effect of Φ on the PCE of various PSCs shown below in

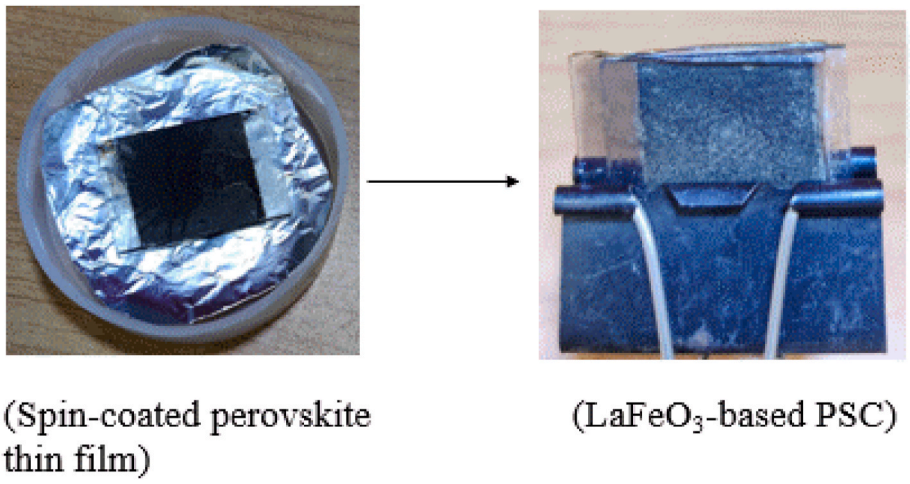


Fig. 23. Fabricated $LaFeO_3$ PSC.

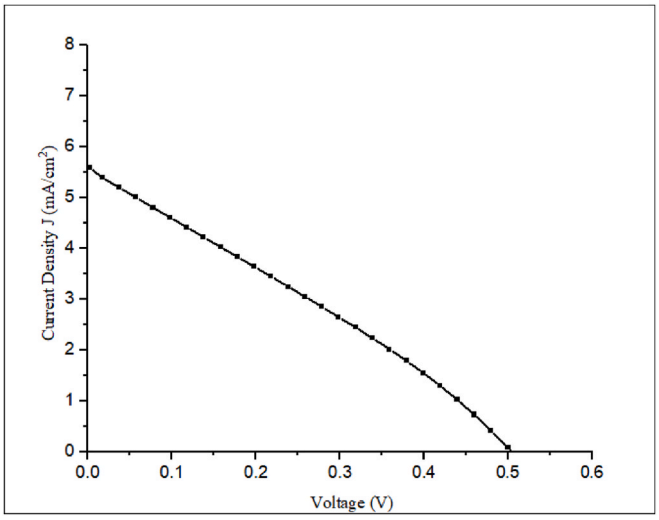


Fig. 24. J-V curve of fabricated $LaFeO_3$-based PSC.

Fig. 20. For the smaller value of work function, performance decreases gradually due to formation of Schottky junction. The enhancement in the PCE with Φ equal to 5.1 eV (Au) was observed. For the large value of work function, PCE remains constant, there is no chemical reaction amid the absorber layer and metal back contacts. The maximum efficiency of 13.03 % achieved at Φ of 5.1 eV (Au).

7. Experimental results comparison with previously reported results

7.1. Simulation results

The PSC structure and energy band diagram of the optimized device demonstrated below in Fig. 21. The optimum efficiency of 13.03 % with V_{OC} = 1.61 V, J_{SC} = 9.952 mA/cm^2, and FF of 81.17 % has been obtained. The optimized ETL and HTL are WS_2 and CBTS respectively with the Au (gold) as metal back contact. The comparison with the previously reported results illustrated in Table 7. The JV and QE curves of optimized PSC are illustrated in Fig. 22.

7.2. Experimental results

The fabricated $LaFeO_3$ PSC and its J-V characteristics are shown below in Figs. 23 and 24 respectively. The PV parameters of fabricated PSC is tabulated below in Table 6. One of the possible reasons for low PCE is the spin-coating process which has certain limitations. That is especially low thermal stability of the film, difficulty in fabricating an ideal multilayer structure, substrate size and shape. Typically, less than 5 % of the initial solution effectively deposits onto the substrate to create the thin film, with the remainder lost due to spinner rotation [57]. The $LaFeO_3$ possesses a bandgap of approximately 2.3 eV [23], which might not be sufficiently wide to prevent rapid recombination of its photogenerated electron-hole pairs which causes low performance of the device. To address this limitation, the preferred approach involves enhancing the E_g through size variation. Decreasing particle size is anticipated to elevate the bandgap, consequently restraining recombination rates. Additionally, optimizing photocatalytic performance involves creating a porous structure to enhance reaction sites. Therefore, augmenting pore density within the particles becomes crucial for maximizing efficiency [23]. $LaFeO_3$-based PSC still cannot be commercialized due to its weak photocatalytic performance, including issues like larger particle size and smaller surface area, reduced crystallinity from low temperature treatment and rapid charge recombination rate [58]. A 10 % doping of transition metal such as Mn, Co and Cu has potential to enhance the photocurrent of $LaFeO_3$ which results in increase in PCE [59].

8. Conclusions

This work comprises synthesis of $LaFeO_3$ perovskite material and simulation of $LaFeO_3$-based PSC followed by experimental validation. The sol-gel method is employed to synthesize $LaFeO_3$ perovskite material. The optimization of ETL and HTL has been done to achieve the maximum PCE. The work deduced that WS_2 as ETL and CBTS as HTL is most appropriate for $LaFeO_3$ PSC. The simulated and experimentally achieved maximum PCE were 13.03 % (with V_{OC} = 1.61 V, J_{SC} = 9.952 mA/cm^2, and FF = 81.17 %) and 0.79 % (with V_{OC} = 0.504 V, J_{SC} = 5.56 mA/cm^2, and FF = 28.31 %) respectively for the configuration of FTO/WS_2/$LaFeO_3$/CBTS/Au. This study also includes the optimization of various factors such as thickness of perovskite layer, operating temperature, N_t, metal back contact, series and shunt resistance. The optimized thickness of $LaFeO_3$ layer is 300 nm and at operating temperature of 300 K. The FTO/WS_2/$LaFeO_3$/CBTS/Au is the most stable structure. The optimized metal back contact is Gold (Au). These outcomes allude that the study is useful to fabricate the high performance $LaFeO_3$-based PSC.

CRediT authorship contribution statement

Rahul Kundara: Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sarita Baghel:** Writing – review & editing, Validation, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The author, Rahul Kundara is thankful to Delhi Technological University, New Delhi for providing a junior research fellowship. We are also very grateful to Marc Burgelman, the Electronics and Information Systems (ELIS), University of Gent, Belgium, for providing us with SCAPS-1D software.

References

[1] S.M. Abdulrahim, Z. Ahmad, M.Q. Mehmood, S. Paek, J. Bhadra, N.J. Al-Thani, M. K. Nazeeruddin, A. Belaidi, M. Amani, Effect of illumination and applied potential on the electrochemical impedance spectra in triple cation (FA/MA/Cs) 3D and 2D/3D perovskite solar cells, J. Electroanal. Chem. 902 (2021) 115800, https://doi.org/10.1016/j.jelechem.2021.115800.

[2] Z. Ahmad, A. Mishra, S.M. Abdulrahim, F. Touati, Electrical equivalent circuit (EEC) based impedance spectroscopy analysis of HTM free perovskite solar cells, J. Electroanal. Chem. 871 (2020) 114294, https://doi.org/10.1016/j.jelechem.2020.114294.

[3] K. Fu, Y. He, B. Zhang, X. Gao, G. Zou, Enhanced aqueous stability and radiative-charge-transfer of CsPbBr3/Ag2S perovskite nanocrystal hybrids, J. Electroanal. Chem. 858 (2020) 113835, https://doi.org/10.1016/j.jelechem.2020.113835.

[4] A. Alizadeh, Z. Shariatinia, Unveiling the influence of SmFeO3-TiO2 nanocomposites as high performance photoanodes of dye-sensitized solar cells, J. Mol. Liq. 348 (2022) 118070, https://doi.org/10.1016/j.molliq.2021.118070.

[5] N. Rai, S. Rai, P.K. Singh, P. Lohia, D.K. Dwivedi, Analysis of various ETL materials for an efficient perovskite solar cell by numerical simulation, J. Mater. Sci. Mater. Electron. 31 (2020) 16269–16280, https://doi.org/10.1007/s10854-020-04175-z.

[6] F. Moradi, Z. Shariatinia, N. Safari, E. Mohajerani, Boosted performances of mesoscopic perovskite solar cells using LaFeO3 inorganic perovskite nanomaterial, J. Electroanal. Chem. 916 (2022) 116376, https://doi.org/10.1016/j.jelechem.2022.116376.

[7] X. Fan, Advanced progress in metal halide perovskite solar cells: a review, Mater, Today Sustain. 24 (2023) 100603, https://doi.org/10.1016/j.mtsust.2023.100603.

[8] K. Domanski, E.A. Alharbi, A. Hagfeldt, M. Grätzel, W. Tress, Systematic investigation of the impact of operation conditions on the degradation behaviour of perovskite solar cells, Nat. Energy 3 (2018) 61–67, https://doi.org/10.1038/s41560-017-0060-5.

[9] D. Yang, R. Yang, S. Priya, S. Liu, Recent advances in flexible perovskite solar cells: fabrication and applications, Angew. Chem. Int. Ed. 58 (2019) 4466–4483, https://doi.org/10.1002/anie.201809781.

[10] R. Kour, S. Arya, S. Verma, J. Gupta, P. Bandhoria, V. Bharti, R. Datt, V. Gupta, Potential substitutes for replacement of lead in perovskite solar cells: a review, Glob, Chall. 3 (2019) 1900050, https://doi.org/10.1002/gch2.201900050.

[11] N.K. Noel, S.D. Stranks, A. Abate, C. Wehrenfennig, S. Guarnera, A.A. Haghighirad, A. Sadhanala, G.E. Eperon, S.K. Pathak, M.B. Johnston, A. Petrozza, Lead-free organic–inorganic tin halide perovskites for photovoltaic applications, Energy Environ. Sci. 7 (2014) 3061–3068, https://doi.org/10.1039/C4EE01076K.

[12] M.G. Ju, M. Chen, Y. Zhou, H.F. Garces, J. Dai, L. Ma, N.P. Padture, X.C. Zeng, Earth-abundant nontoxic titanium (IV)-based vacancy-ordered double perovskite halides with tunable 1.0 to 1.8 eV bandgaps for photovoltaic applications, ACS Energy Lett. 3 (2018) 297–304, https://doi.org/10.1021/acsenergylett.7b01167.

[13] Z. Zhang, X. Li, X. Xia, Z. Wang, Z. Huang, B. Lei, Y. Gao, High-quality (CH3NH3)3Bi2I9 film-based solar cells: pushing efficiency up to 1.64, J. Phys. Chem. Lett. 8 (2017) 4300–4307, https://doi.org/10.1021/acs.jpclett.7b01952.

[14] M. Wang, P. Zeng, S. Bai, J. Gu, F. Li, Z. Yang, M. Liu, High-quality sequential-vapor-deposited Cs2AgBiBr6 thin films for lead-free perovskite solar cells, Sol. RRL 2 (2018) 1800217, https://doi.org/10.1002/solr.201800217.

[15] D. Çoban Özkan, A. Türk, E. Çelik, Synthesis and characterizations of sol–gel derived LaFeO3 perovskite powders, J. Mater. Sci. Mater. Electron. 31 (2020) 22789–22809, https://doi.org/10.1007/s10854-020-04803-8.

[16] K.R. Nagavenkatesh, C. Sambathkumar, M. Murugesan, P. Devendran, N. Nallamuthu, P. Lakshmanan, A. Shameem, Review on the recent improvements in lanthanum ferrite perovskites for visible light driven textile dyes degradation and its various hetero-junctions catalysis mechanism, J. Ind. Eng. Chem. (2024), https://doi.org/10.1016/j.jiec.2024.02.024.

[17] S.L. Yu, M. Wang, E. Ye, S. Lei, J. Liu, G. Liu, G. Qiao, Synergistic enhancement of optical properties in Ca/Ni co-doped LaFeO3 perovskite coatings enabling high-temperature photothermal conversion, Appl. Surf. Sci. 642 (2024) 158595, https://doi.org/10.1016/j.apsusc.2023.158595.

[18] M. Sabir, N. AlMasoud, M. Ramzan, M. Aamir, S.R. Ejaz, T.S. Alomar, Z.M. El-Bahy, M.A. Salam, S.M. Albukhari, D.F. Baamer, Rare earth and transition metal co-doped LaFeO3 perovskite and its CNTs reinforced nanohybrid for environmental remediation application, Ceram. Int. 49 (2023) 20939–20950, https://doi.org/10.1016/j.ceramint.2023.03.227.

[19] S. Bai, B. Shi, L. Ma, P. Yang, Z. Liu, D. Li, A. Chen, Synthesis of LaFeO3 catalytic materials and their sensing properties, Sci. China, Ser. B: Chem. 52 (2009) 2106–2113, https://doi.org/10.1007/s11426-009-0289-3.

[20] M.K. Das, S. Panda, N. Mohapatra, Power conversion efficiency optimization of LaFeO3 Mott insulator based solar cell with metal oxide transport layers using SCAPS, Mater Today Proc 74 (2023) 756–762, https://doi.org/10.1016/j.matpr.2022.11.031.

[21] L.N. Moghadam, Z.R. Ranjbar, Cost-efficient solar cells using nanocrystalline perovskite La (Fe and Mn) O3 and candle soot: theory and experiment, J. Alloys Compd. 785 (2019) 117–124, https://doi.org/10.1016/j.jallcom.2019.01.068.

[22] R. Chakraborty, K.M. Sim, M. Shrivastava, K.V. Adarsh, D.S. Chung, A. Nag, Colloidal synthesis, optical properties, and hole transport layer applications of Cu2BaSnS4 (CBTS) nanocrystals, ACS Appl. Energy Mater. 2 (2019) 3049–3055, https://doi.org/10.1021/acsaem.9b00473.

[23] W. Azouzi, W. Sigle, H. Labrim, M. Benaissa, Sol-gel synthesis of nanoporous LaFeO3 powders for solar applications, Mater. Sci. Semicond. Process. 104 (2019) 104682, https://doi.org/10.1016/j.mssp.2019.104682.

[24] W. Zheng, R. Liu, D. Peng, G. Meng, Hydrothermal synthesis of LaFeO3 under carbonate-containing medium, Mater. Lett. 43 (2000) 19–22, https://doi.org/10.1016/S0167-577X(99)00223-2.

[25] M. Romero, R.W. Gómez, V. Marquina, J.L. Pérez-Mazariego, R. Escamilla, Synthesis by molten salt method of the AFeO3 system (A=La, Gd) and its structural, vibrational and internal hyperfine magnetic field characterization, Physica B Condens Matter 443 (2014) 90–94, https://doi.org/10.1016/j.physb.2014.03.024.

[26] N. Sharma, H.S. Kushwaha, S.K. Sharma, K. Sachdev, Fabrication of LaFeO3 and rGO-LaFeO3 microspheres based gas sensors for detection of NO2 and CO, RSC Adv. 10 (2020) 1297–1308, https://doi.org/10.1039/c9ra09460a.

[27] X. Ren, H. Yang, S. Gen, J. Zhou, T. Yang, X. Zhang, Z. Cheng, S. Sun, Controlled growth of LaFeO3 nanoparticles on reduced graphene oxide for highly efficient photocatalysis, Nanoscale 8 (2016) 752–756, https://doi.org/10.1039/c5nr06338h.

[28] H. Bencherif, L. Dehimi, N. Mahsar, E. Kouriche, F. Pezzimenti, Modeling and optimization of CZTS kesterite solar cells using TiO2 as efficient electron transport layer, Mater. Sci. Eng. B. 276 (2022) 115574, https://doi.org/10.1016/j.mseb.2021.115574.

[29] A.A. Kanoun, M.B. Kanoun, A.E. Merad, S. Goumri-Said, Toward development of high-performance perovskite solar cells based on CH3NH3GeI3 using computational approach, Sol. Energy 182 (2019) 237–244, https://doi.org/10.1016/j.solener.2019.02.041.

[30] K. Deepthi Jayan, V. Sebastian, Comprehensive device modelling and performance analysis of MASnI3 based perovskite solar cells with diverse ETM, HTM and back metal contacts, Sol. Energy 217 (2021) 40–48, https://doi.org/10.1016/j.solener.2021.01.058.

[31] A. Raj, M. Kumar, H. Bherwani, A. Gupta, A. Anshul, Evidence of improved power conversion efficiency in lead-free CsGeI 3 based perovskite solar cell heterostructure via scaps simulation, J. Vac. Sci. Technol. B. 39 (2021) 012401, https://doi.org/10.1116/6.0000718.

[32] A. Husainat, W. Ali, P. Cofie, J. Attia, J. Fuller, Simulation and Analysis of Methylammonium Lead Iodide (CH3NH3PbI3) Perovskite Solar Cell with Au Contact Using SCAPS 1D Simulator, vol. 7, American Journal of Optics and Photonics, 2019, pp. 33–40, https://doi.org/10.11648/j.ajop.20190702.12.

[33] S. Bhattarai, T.D. Das, Optimization of carrier transport materials for the performance enhancement of the MAGeI3 based perovskite solar cell, Sol. Energy 217 (2021) 200–207, https://doi.org/10.1016/j.solener.2021.02.002.

[34] H. Bencherif, M.K. Hossain, Design and numerical investigation of efficient (FAPbI3) 1− x (CsSnI3) x perovskite solar cell with optimized performances, Sol. Energy 248 (2022) 137–148, https://doi.org/10.1016/j.solener.2022.11.012.

[35] U. Mandadapu, Simulation and analysis of lead based perovskite solar cell using SCAPS-1D, Indian J. Sci. Technol. 10 (2017) 1–8, https://doi.org/10.17485/ijst/2017/v11i10/110721.

[36] F. Azri, A. Meftah, N. Sengouga, A. Meftah, Electron and hole transport layers optimization by numerical simulation of a perovskite solar cell, Sol. Energy 181 (2019) 372–378, https://doi.org/10.1016/j.solener.2019.02.017.

[37] S. Mushtaq, S. Tahir, A. Ashfaq, R. Sebastian Bonilla, M. Haneef, R. Saeed, W. Ahmad, N. Amin, Performance optimization of lead-free MASnBr3 based perovskite solar cells by SCAPS-1D device simulation, Sol. Energy 249 (2023) 401–413, https://doi.org/10.1016/j.solener.2022.11.050.

[38] M.K. Hossain, G.F.I. Toki, A. Kuddus, M.H.K. Rubel, M.M. Hossain, H. Bencherif, M.F. Rahman, M.R. Islam, M. Mushtaq, An extensive study on multiple ETL and HTL layers to design and simulation of high-performance lead-free CsSnCl3-based

perovskite solar cells, Sci. Rep. 13 (2023) 2521, https://doi.org/10.1038/s41598-023-28506-2.
[39] N. Lakhdar, A. Hima, Electron transport material effect on performance of perovskite solar cells based on CH3NH3GeI3, Opt. Mater. 99 (2020) 109517, https://doi.org/10.1016/j.optmat.2019.109517.
[40] M. Singh, R. Singh, V. Kumar, V. Singh, Investigation of CH3NH3PbI3 and CH3NH3SnI3 based perovskite solar cells with CuInSe2 nanocrystals, Optik 246 (2021) 167839, https://doi.org/10.1016/j.ijleo.2021.167839.
[41] I. Montoya De Los Santos, H.J. Cortina-Marrero, M.A. Ruíz-Sánchez, L. Hechavarría-Difur, F.J. Sánchez-Rodríguez, M. Courel, H. Hu, Optimization of CH3NH3PbI3 perovskite solar cells: a theoretical and experimental study, Sol. Energy 199 (2020) 198–205, https://doi.org/10.1016/j.solener.2020.02.026.
[42] H. Bencherif, F. Meddour, M.H. Elshorbagy, M.K. Hossain, A. Cuadrado, M.A. Abdi, T. Bendib, S. Kouda, J. Alda, Performance enhancement of (FAPbI3) 1-x (MAPbBr3) x perovskite solar cell with an optimized design, Micro Nanostructures 171 (2022) 207403, https://doi.org/10.1016/j.micrna.2022.207403.
[43] A. Kuddus, M.F. Rahman, S. Ahmmed, J. Hossain, A.B.M. Ismail, Role of facile synthesized V2O5 as hole transport layer for CdS/CdTe heterojunction solar cell: validation of simulation using experimental data, Superlattices Microstruct 132 (2019) 106168, https://doi.org/10.1016/j.spmi.2019.106168.
[44] M.K. Hossain, A.A. Arnab, R.C. Das, K.M. Hossain, M.H.K. Rubel, M.F. Rahman, H. Bencherif, M.E. Emetere, M.K.A. Mohammed, R. Pandey, Combined DFT, SCAPS-1D, and wxAMPS frameworks for design optimization of efficient Cs2BiAgI6-based perovskite solar cells with different charge transport layers, RSC Adv. 12 (2022) 35002–35025, https://doi.org/10.1039/d2ra06734j.
[45] S. Ahmed, F. Jannat, M.A.K. Khan, M.A. Alim, Numerical development of eco-friendly Cs2TiBr6 based perovskite solar cell with all-inorganic charge transport materials via SCAPS-1D, Optik 225 (2021) 165765, https://doi.org/10.1016/j.ijleo.2020.165765.
[46] S. Rai, B.K. Pandey, D.K. Dwivedi, Modeling of highly efficient and low cost CH3NH3Pb(I1-xClx)3 based perovskite solar cell by numerical simulation, Opt. Mater. 100 (2020) 109631, https://doi.org/10.1016/j.optmat.2019.109631.
[47] M.B. Johnston, L.M. Herz, Hybrid perovskites for photovoltaics: charge-carrier recombination, diffusion, and radiative efficiencies, Acc. Chem. Res. 49 (2016) 146–154, https://doi.org/10.1021/acs.accounts.5b00411.
[48] H. Bencherif, Towards a high efficient Cd-free double CZTS layers kesterite solar cell using an optimized interface band alignment, Sol. Energy 238 (2022) 114–125, https://doi.org/10.1016/j.solener.2022.04.040.
[49] M.S. Islam, K. Sobayel, A. Al-Kahtani, M.A. Islam, G. Muhammad, N. Amin, M. Shahiduzzaman, M. Akhtaruzzaman, Defect study and modelling of SnX3-based perovskite solar cells with SCAPS-1D, Nanomaterials 11 (2021) 1218, https://doi.org/10.3390/nano11051218.
[50] K. Poorkazem, D. Liu, T.L. Kelly, Fatigue resistance of a flexible, efficient, and metal oxide-free perovskite solar cell, J Mater Chem A Mater 3 (2015) 9241–9248, https://doi.org/10.1039/c5ta00084j.
[51] E. Danladi, A.C. Egbugha, R.C. Obasi, N.N. Tasie, C.U. Achem Cu, I.S. Haruna, L. O. Ezeh, Defect and doping concentration study with series and shunt resistance influence on graphene modified perovskite solar cell: a numerical investigation in SCAPS-1D framework, J. Indian Chem. Soc. 100 (2023) 101001, https://doi.org/10.1016/j.jics.2023.101001.
[52] S. Karthick, S. Velumani, J. Bouclé, Experimental and SCAPS simulated formamidinium perovskite solar cells: a comparison of device performance, Sol. Energy 205 (2020) 349–357, https://doi.org/10.1016/j.solener.2020.05.041.
[53] H. Heriche, Z. Rouabah, N. Bouarissa, New ultra thin CIGS structure solar cells using SCAPS simulation program, Int. J. Hydrogen Energy 42 (2017) 9524–9532, https://doi.org/10.1016/j.ijhydene.2017.02.099.
[54] A.K. Sharma, N.K. Chourasia, P.K. Jha, R. Kumar, M. Kumar, R.K. Chourasia, Characteristic features and performance investigations of a PTB7:PC71BM/PFN:Br pure organic solar cell using SCAPS-1D, J. Electron. Mater. 52 (2023) 4302–4311, https://doi.org/10.1007/s11664-022-10202-y.
[55] P.K. Jha, N.K. Chourasia, A.K. Sharma, R.K. Chourasia, Optimization of electrical properties for performance analysis of p-Si/n-CdS/ITO heterojunction photovoltaic cell, Mater Today Proc 67 (2022) 620–624, https://doi.org/10.1016/j.matpr.2022.05.577.
[56] G. Pindolia, S.M. Shinde, P.K. Jha, Non-leaded, KSnI3 based perovskite solar cell: a DFT study along with SCAPS simulation, Mater. Chem. Phys. 297 (2023) 127426, https://doi.org/10.1016/j.matchemphys.2023.127426.
[57] R. Nistico, D. Scalarone, G. Magnacca, Sol-gel chemistry, templating and spin-coating deposition: a combined approach to control in a simple way the porosity of inorganic thin films/coatings, Micropor Mesopor Mat 248 (2017) 18–29, https://doi.org/10.1016/j.micromeso.2017.04.017.
[58] M. Humayun, H. Ullah, M. Usman, A. Habibi-Yangjeh, A.A. Tahir, C. Wang, W. Luo, Perovskite-type lanthanum ferrite based photocatalysts: preparation, properties, and applications, J. Energy Chem. 66 (2022) 314–338, https://doi.org/10.1016/j.jechem.2021.08.023.
[59] Q. Peng, B. Shan, Y. Wen, R. Chen, Enhanced charge transport of LaFeO3 via transition metal (Mn, Co, Cu) doping for visible light photoelectrochemical water oxidation, Int. J. Hydrog. Energy. 40 (2015) 15423–15431, https://doi.org/10.1016/j.ijhydene.2015.09.072.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 May 2024, accepted 27 May 2024, date of publication 4 June 2024, date of current version 12 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3409424

RESEARCH ARTICLE

Performance Assessment of Underwater-to-Air Optical Wireless Communication System With the Effect of Solar Noise and Sea Surface Conditions

BHOGESWARA RAO ANGARA[1,2], PALANISAMY SHANMUGAM[1], HARISHANKAR RAMACHANDRAN[2], AND CHAVAPATI GOUSE SANDHANI[1]

[1]Department of Ocean Engineering, Indian Institute of Technology Madras, Chennai 600036, India
[2]Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai 600036, India

Corresponding author: Palanisamy Shanmugam (pshanmugam@iitm.ac.in)

This work was supported by the Ministry of Human Resource Development (MHRD) Fellowship through the Interdisciplinary Research Program (IDRP) with the Indian Institute of Technology (IIT) Madras.

ABSTRACT This study focuses on the performance evaluation of underwater-to-air Optical Wireless Communication (OWC) and data transmission between Autonomous Underwater Vehicles (AUVs) and Autonomous Aerial Vehicles (AAVs). Facing challenges from the harsh marine environment and optical signal attenuation, we utilized Monte Carlo (MC) simulations to comprehensively assess the system performance. Our investigation incorporated crucial elements impacting signal attenuation, including particulate matter, bubbles, rough sea conditions, and solar noise. By integrating the in-situ Inherent Optical Properties (IOPs) data and Hall-Novarini (HN) model for bubble characteristics, we constructed a robust simulation framework. To accurately estimate optical beam fluctuations caused by surface waves, we employed the three-dimensional theoretical ECKV (Elfouhaily, Chapron, Katsaros, and Vandemark) model for the water-air interface layer. The received power calculated for varying wind conditions, with and without bubbles, revealed fluctuations attributed to surface waves' interaction with the optical signal. Moreover, we observed a decrease in received power due to the bubble effects and the water-air interface layer. Solar radiation's significant impact on the transmitted signal, especially at the airborne receivers, was evident through the BER performance analysis. We conducted extensive experiments for measuring the solar radiance and evaluating its influence on the BER performance, particularly noting degradation during zenith sun positions and increased errors in scenarios with higher solar glint. Additionally, our investigation highlighted the adverse effect of sea surface bubbles on received solar power and the findings revealed that the BER performance is significantly reduced in the presence of bubbles rather than the scenarios without bubbles.

INDEX TERMS Optical wireless communication, Monte Carlo, optical properties, bubbles, solar noise, oceanic water.

I. INTRODUCTION

Underwater-to-air Optical Wireless Communication (OWC) is a rapidly growing field because of its significant impacts across a range of industries. It enables real-time data transfer from the submerged source to the surface systems, supporting scientific research, environmental monitoring, national security, and industrial/maritime applications. Underwater Optical Wireless Communications (UOWC) offer high bandwidth, making it a foundational tool for diverse applications due to the weak absorption of the water in the blue-green wavelength region [1], [2]. Traditional methods

The associate editor coordinating the review of this manuscript and approving it for publication was Yougan Chen.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

like acoustic [3], [4] and radio waves [5], [6] have limitations in data rate and propagation. There is a growing interest in exploring and exploiting marine natural resources, which require an improved UOWC system for communication within/between the underwater sensor networks, autonomous underwater vehicles (AUVs), remotely-operated vehicles (ROVs), and from submarines to autonomous aerial vehicles (AAVs). UOWC is a promising technology for such applications, offering high data rates, low latency, and energy-efficient communication links [7], particularly in clear waters with distances of up to approximately 150 m [8], [9].

Earlier researchers have conducted comprehensive studies on modelling the downlink (air-to-underwater) and underwater-to-underwater OWC channel in various ocean waters. When modelling the downlink UOWC channel, researchers employed the Cox and Munk method to model the sea surface. Nevertheless, this method lacks essential spatial and temporal correlated information necessary for accurately modelling the channel for laser communication. Furthermore, their investigations were carried out without considering solar noise. However, as of now, there have been no studies conducted on modeling the underwater-to-air OWC channel. The study of underwater-to-air OWC systems is of paramount importance due to the increasing demand for real-time applications in underwater environments. In scenarios such as underwater sensor networks and AUVs, the need for timely and efficient transmission of data from the underwater devices to the surface is crucial. This underwater-to-air communication plays a vital role in applications where instant data collection and monitoring are required, driving the focus on enhancing the reliability and efficiency of the uplink UOWC systems to meet the growing demands of real-time underwater communication.

In underwater-to-air OWC systems, there is a significant challenge posed by an optical pathway transition from the underwater environment to the above-surface air medium. The wavy ocean surface produced by wind fields represents one of the major constraints that impact the reliability and performance of communication links. Wind-induced waves, currents, and other environmental factors create disturbances in the water-air interface layer, which can lead to fluctuations in the refractive index gradient at the water-air interface and cause scattering, diffraction, and turbulence in the optical signal. Researchers have developed various models to describe the air-sea interface and its impact on optical communication. These models aim to simulate and understand the behaviour of light as it traverses through this complex boundary. Some of the commonly used sea surface models include the Boussinesq and Korteweg-de Vries (KdV) model [10], Cox and Munk model [11] and Fast Fourier Transform (FFT) approximation [12]. The air-sea interface models find extensive applications, particularly in remote sensing. They play a crucial role in estimating various parameters, including bidirectional transmission and reflectance distribution functions (BTDF and BRDF) [13], [14]. These models have proven useful in understanding how the reflected light from the sea surface impacts our observations, especially when conditions such as sky radiance and sun glint are present. Further, these models have been used to analyse the water-leaving radiance in different atmospheric and marine conditions [15], [16]. Earlier studies have relayed on the probabilistic sea surface models, originally developed by Cox and Munk, to study the effect of the air-sea interface [14], [17], [18]. Numerous studies have utilised the Cox and Munk model to analyse UOWC vertical links, particularly when the light beam propagates from the air medium (atmosphere) to the underwater environment. These studies observed that the greater variability in sea surface slopes due to wind speed (u) reduces the system performance, which resulted in higher power loss and increased bit error rates [19], [20], [21]. Studies have also utilised the KdV equations to simulate the air-sea interface with an aim to analyse the coverage area of an optical beam as it propagates from the air into the underwater environment, considering the presence of multiple light sources located in the air medium [22]. In a controlled laboratory setting, experiments were conducted to study a short-range (0.66 m) water-to-air OWC link. These experiments indicated fluctuations in the channel link gain under both calm and wavy surface conditions [23]. In recent studies [24], researchers have developed a model to simulate the vertical UOWC channel. This model takes into account several factors simultaneously: oceanic turbulence parameters, real-time depth-dependent temperature and salinity, chlorophyll concentration, and the presence of bubbles. Additionally, it considers the transmitter beam's oblique incidence at various angles. The findings indicate that the scintillation index rises as the transmitter tilt angle increases. The study [25] evaluates UOWC links with on-off key modulation and selection combining (SC) receive diversity under underwater conditions. It derives closed-form expressions for BER and outage probability, validated through simulations, showing SC diversity mitigates beam attenuation effects. The study [26] examines vertical UOWC links affected by non-mixing layers due to ocean stratification, causing turbulence variations. It models the channel as a concatenation of these layers, deriving closed-form expressions for BER and analysing diversity gains. Results reveal inaccuracies in assuming constant turbulence strength and provide insights into the effect of layering on capacity. Another study [27] explores the performance of vertical UWOC systems amidst air bubbles and temperature gradients. A layered structure with N layers is employed to model the vertical channel's turbulence inhomogeneity using an exponential generalised gamma distribution. Asymptotic average BER and ergodic capacity are derived using H-functions and validated through simulation results. In our most recent investigation [28], we thoroughly examined how sea surface waves and bubbles affect underwater-to-air OWC systems. Our analysis

indicates a substantial variation in received power, depending on the wind speed above the sea surface.

Another challenge in modelling the underwater-to-air OWC channel is the presence of oceanic bubbles. These bubbles are generated by breaking waves in the upper ocean [29], [30] and penetrate the lower water column at depths over several meters, strongly contingent on wind speeds [31]. As a consequence of these bubbles in the upper ocean layer, the light beam is randomly affected and attenuated when passing through the bubble layer. Under high wind speeds, bubble scattering is rapidly enhanced and even more pronounced than particulate scattering in the air-water interface layer. Thus, it is crucial to understand the effect of bubbles on the performance of the underwater-to-air OWC system. Numerous studies revealed a substantial degradation in UOWC system performance in the presence of bubbles (including clean-coated variants) of varying sizes under different wind speeds [19], [32].

Another significant effect on the UOWC system is the interference of solar radiation with an optical signal received by a photodetector. This interference is particularly more pronounced when the receiver is situated in the atmosphere/air medium rather than in the underwater environment. The spectrum of sunlight covers the entire visible wavelength range and beyond in clear marine environments with high water transparency, and it can penetrate several meters of water. Particularly, longer blue to shorter green wavelengths of light can penetrate deeper into the water due to low absorption of water, which allows the penetrating radiation to be easily detectable by sensitive detectors even at depths exceeding 600 m [33]. Previous studies have not paid much attention to the influence of solar noise on the UOWC performance. However, recent work has highlighted background noise's impact and how it affects the UOWC system through various factors like the link length, receiver field of view, and different types of photodetectors [34], [35], [36], [37], [38]. In [34], the significant impact of solar light interference on the UOWC system performance was focused, and effective noise cancellation techniques were emphasised to optimise the UOWC systems in high-intensity solar light conditions. This domain becomes more complex when the receiver is located in the air medium, where it is exposed to solar noise influenced by several factors, including sky radiance, sea surface roughness, water-leaving radiance (depending on the wind speed and wave shape), the time of day, and sun glint.

In this study, our main focus is to incorporate the above factors into the channel model and investigate the performance of underwater-to-air OWC system under different ocean and surface conditions. The Monte-Carlo (MC) technique is employed to simulate the underwater-to-air OWC system with in-situ optical data collected from oceanic waters in the Bay of Bengal and Southern Ocean. Ocean bubble characteristics are determined using the widely recognised Hall-Novarini (HN) model [39], and the distribution of bubble sizes and their ability to penetrate deeper into water depends upon the prevailing wind speed. Recognising the limitations of the Cox and Munk sea surface model, which lacks spatial and temporal correlation information, we utilise the ECKV sea surface model because of its ability to incorporate both spatial and temporal correlation data. This information is essential in understanding how the light beam propagates through the air-sea interface. We assess the solar interference experienced by the receiver by analysing the radiometric data collected above and below the water surface. These data were collected at various times of the day and under different sea conditions, including calm waters, wavy surfaces, and conditions with wind-induced bubbles. Our analysis reveals that the amount of solar power captured by the receiver varies significantly depending on these parameters.

In addition, we examine the correlation between received power, wind speed, and wind-induced bubbles in an underwater-to-air OWC system. The results are compared across a range of sea surface conditions. Finally, the BER performance of underwater-to-air OWC systems was assessed, considering the individual and combined effects of waves, wind-induced bubbles, and varying levels of solar interference.

II. UNDERWATER-TO-AIR OWC SYSTEM MODEL

This section presents the concepts of an underwater-to-air OWC system under investigation (Fig. 1) and describes the key components of the models considered in this study. Here, the transmitter (laser source) is positioned at a specific distance beneath the sea surface, while the receiver is situated several meters above the sea surface. As the optical signal propagated from the source to the destination, it can undergo attenuation due to the various phenomena depicted in Fig. 1. Thus, it becomes crucial to comprehend the underlying theories related to sea surface models, solar noise models, the mechanisms involving bubbles in seawater, and the inherent optical properties (IOPs) of seawater. The fundamental IOPs include absorption (a), scattering (b), and attenuation ($c = a + b$) coefficients and scattering phase function, which collectively play a pivotal role in characterising the optical behaviour of seawater. These parameters were employed as the essential inputs in our MC method for effectively simulating the propagation of photons within the water column.

A. SEA SURFACE GENERATION USING ECKV MODEL

The two-dimensional ECKV (Elfouhaily, Chapron, Katsaros, and Vandemark) spectrum serves as a fundamental tool in oceanography and sea surface modelling [12]. It provides a spectral description of wave energy distribution, which is used to generate realistic sea surfaces through a combination of mathematical operations and random phase introduction. Its applications extend beyond sea surface generation to include areas like remote sensing, environmental impact assessment, and the study of nonlinear wave effects. To obtain the sea surface elevation $z\left(x_i, y_j\right)$, the points on the sea surface grid are sampled in a region of size L_x by L_y m with N_x and N_y points along x and y-directions, respectively,

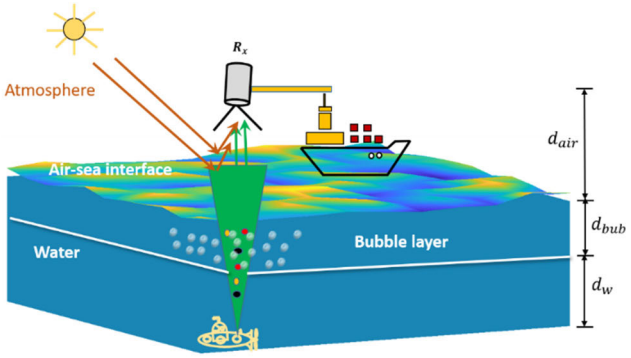


FIGURE 1. Schematic of the channel model for the underwater-to-air OWC system.

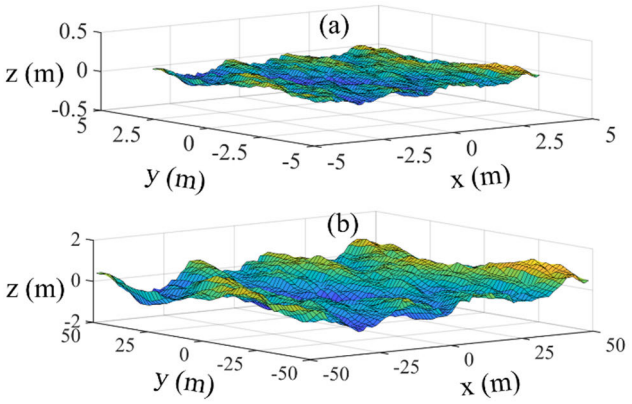


FIGURE 2. The sea surface realisations using the ECKV model for a wind speed of a) 3 m s^{-1} b) 10 m s^{-1}.

as given by

$$z\left(x_i, y_j\right) = \sum_{u=o}^{N_x-1} \sum_{v=0}^{N_y-1} \hat{z}(u, v) e^{i2\pi(x_i u/N_x + y_j v/N_y)} \quad (1)$$

where the harmonic components $\hat{z}(u, v)$ are indexed by u and v in the wavenumber k domain. For an understanding of the ECKV sea surface model, the theoretical foundations and derivations of this model can be found in a previous study [12].

The process of generating a sea surface using the ECKV spectrum involves several key steps. First, it requires the computation of the two-sided discrete values of the elevation variance, which essentially quantifies how wave energy is distributed across different frequencies and directions in the ocean. Next, in order to create a realistic sea surface, the simulation employs random sampling techniques. Specifically, it samples the amplitudes of the harmonic components from normal distributions. This introduces variability and randomness into the wave field while preserving the spectral characteristics defined by the ECKV spectrum. Finally, to translate the spectral information into the spatial domain and obtain the sea-surface elevation in two dimensions, the Inverse Fast Fourier Transform (IFFT) is employed. This mathematical operation is essential for generating a detailed representation of the sea surface, complete wave crests, troughs, and their variations across space and time. These three interconnected steps collectively contribute to the creation of a realistic and dynamic sea surface based on the ECKV spectrum.

The ECKV model originates from the linear superposition theory and excels in accurately representing the complex dynamics of the ocean's surface. The ECKV model integrates both spatial (between adjacent triangle vertices) and temporal (from one time step to the next) correlation information into the generated sea surface model. Thus, it is crucial to understand the spatial and temporal correlation of sea surface data in order to gain insights into the propagation of laser beams across the water-air interface. The sea surfaces generated using the ECKV method, with wind speeds of 3 and 10 m s^{-1}, are depicted in Fig. 2. As the wind speed increases, the sea surface becomes progressively rougher. Consequently, the mean square slope value, which quantifies the degree of surface roughness, increases in direct correlation with higher wind speeds.

B. PHOTON PROPAGATION THROUGH THE WATER, WATER-AIR INTERFACE AND AIR USING MONTE-CARLO METHOD

The MC approach is invaluable for modelling and understanding the complex interactions between light and water, finding diverse applications that extend to areas like underwater communication. Figure 3 illustrates the key steps of Monte Carlo simulations used to model the propagation of photons from underwater to air medium. In the MC simulation method for photon propagation in a water column, several key steps are involved. First, the optical properties of water, including the absorption (a), scattering (b) and attenuation (c) coefficients, are defined. Next, the photon's initial position and direction are determined. The simulation then progresses through discrete steps, with the photon's trajectory being probabilistically determined based on these properties. At each step, the photon's interactions with the medium are governed by the scattering or absorption events, contingent upon the scattering albedo ($\omega = b/c$) condition. The probabilities for these interactions are calculated using the relevant optical properties. When the photon moves within the water column, whether it is absorbed or scattered is determined by comparing a randomly generated number ξ, drawn from a uniform distribution [0 1], with the albedo value (ω). If ξ is less than or equal to ω, the photon undergoes scattering; otherwise, it is absorbed by the water particles. In instances where a photon is scattered by a water particle or bubble, the scattering angle is determined by referencing the respective volume scattering function (VSF). A detailed discussion about the MC rules for photon propagation through the water medium and air-sea interface is given in [19] and [21]. In these simulations, we generated sea surface models using the ECKV method, which relies on the FFT technique. The sea surface is generated with an optimal selection of

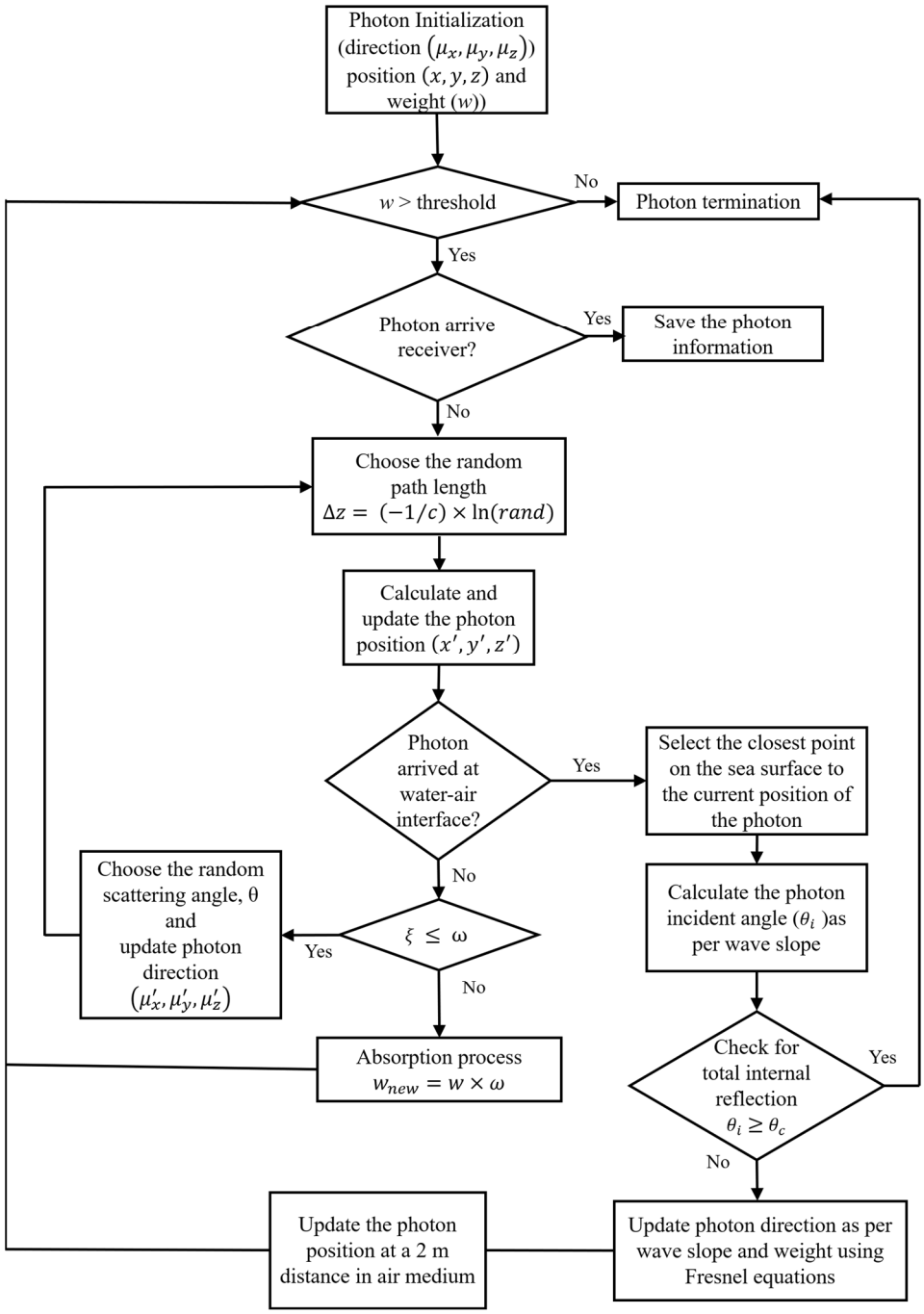


FIGURE 3. A flowchart of photon propagation from the underwater-to-air OWC system.

spatial grid resolution, a choice that depends on both wind speed and maximum wavelength of the wave.

During the photon's propagation, once it has traversed a distance of (d_w+ d_{bub}) m within the water and bubble layer and reached the sea surface, the simulation identifies the point on the sea surface that is closest to the photon's current position. The photon's position is then updated accordingly, with the simulation choosing the point on the sea surface that offers the minimum distance from the photon's current location. To compute the normal vectors ($\hat{n}$) of the sea surface at individual grid points, we employ bicubic interpolation. This involves first applying a bicubic data fitting process to grid data. Subsequently, we derive diagonal vectors and employ them to ascertain the surface normal at each vertex through a cross-product operation. We determine the angle at which a photon strikes the wave surface, denoted as (θ_i), by taking the scalar product of two vectors: the unit normal vector ($\hat{n}$) and the direction cosine vector ($\hat{\mu}_i$) representing

the photon's direction just before it reaches the water-air interface.

Subsequently, we calculate the angle of transmittance (θ_t) and the transmission coefficient (ρ) specific to the wave facet. These values are derived using the Snell's law and Fresnel equations [21], which provides essential insights into the photon's behaviour as it interacts with the wave.

As the photon travels a distance of a few meters through the air before reaching the receiver, our working assumption is that the maritime clear atmosphere remains free from particulate matter. This implies that during its propagation, the photon encounters no significant attenuation ($c \approx 0.9 \times 10^{-4}\text{m}^{-1}$) phenomena that would alter its path or intensity.

C. IN-SITU IOPS OF SEAWATER WITH AND WITHOUT BUBBLES

Previous studies have relied on the constant (single) values of absorption and scattering coefficients, typically derived from the Petzold measurements. However, these values are suitable for horizontal UOWC links. For vertical links, it is important to note that the attenuation and scattering profiles are not constant as these coefficients vary with depth due to the particulate matters (including sediments, phytoplankton, detritus, and bubbles concentrated in the upper sea surface layers). These IOPs were determined through the measurements obtained from optical instruments (such as the Wet Lab's ACS and LISST-VSF) in Bay of Bengal waters. These measurements were used as the inputs for MC simulations. The in-situ measurements were collected under normal sea state conditions without much bubbles. Statistical information regarding the bubbles was obtained from the HN model [39]. The IOPs of oceanic bubbles were then calculated using the HN model and methods established in earlier studies [31].

The VSF data, which include the influence of bubbles (following a power law distribution), were also taken from the measurements conducted in Southern Ocean waters [40]. These measurements were collected in clear waters where the concentrations of chlorophyll and total suspended matter were recorded as 0.72 mg m^{-3} and 0.5 g m^{-3} respectively. The scattering phase function data were obtained by normalising the VSF values with the scattering coefficients. The resulting scattering phase functions are shown in Fig. 4, where bubbles exhibit distinct scattering characteristics. The results indicate a higher probability of scattering at larger angles in the presence of bubbles compared to the cases without bubbles.

Figure 5 presents the depth profiles of both total scattering (b_t) and attenuation coefficients (c_t), which were measured in open sea waters of the Bay of Bengal. For this study, we considered a bubble distribution that reflects the prevalent bubble sizes under real-time oceanic conditions. These bubbles follow a power law distribution characterised by the minimum (r_{min}) and maximum bubble sizes (r_{max}) of 5 μm and 200 μm respectively, with a distribution slope (α) of - 4.

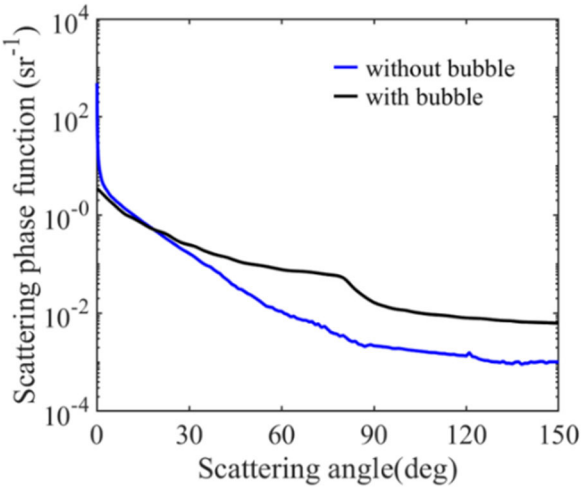


FIGURE 4. Scattering phase functions compared between bubble-free Bay of Bengal waters and bubble-containing Southern Ocean waters.

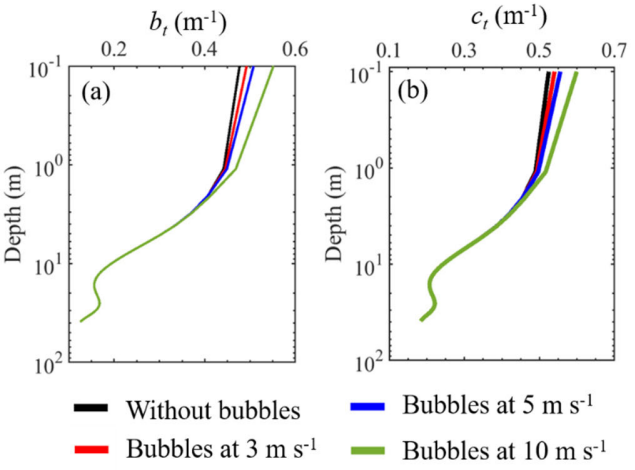


FIGURE 5. Depth profiles of (a) scattering and (b) attenuation coefficients collected in Bay of Bengal waters (with and without bubbles).

Assuming the absorption coefficient of bubbles being negligible, the scattering coefficients of bubbles were calculated using [41]

$$b_{bub}(z) = N(z)\,\bar{Q}_{sca}\bar{S} \tag{2}$$

where N(z) represents the bubble density at a given depth (z) within the water column, which was calculated using the HN model, $\bar{S}$ denotes the geometric mean cross-sectional area of bubbles and $\bar{Q}_{sca}$ represents the mean scattering efficiency.

The measurements indicate that both b_t and c_t values decrease towards the deeper water column as a result of decreasing particulate matter concentrations. Additionally, with an increase in wind speed, there is an increase in the number density of bubbles $N(z)$. Consequently, this leads to an increase in the scattering and attenuation coefficients attributed to bubbles (according to Eq. 2).

D. SOLAR NOISE MODELLING

To quantify the solar radiance received by the receiver (positioned a few meters above the sea surface and looking vertically downward toward the air/water boundary), we evaluated its influence on the performance of the underwater-to-air OWC system by analysing the total upwelling solar radiance spectrum (denoted as $L_t(\theta, \lambda)$ in units of W m^{-2} nm^{-1} sr^{-1}, where λ and θ represent the wavelength and solar zenith angle respectively). Due to the position of receiver in the air medium, the optical signal of reflected solar radiation (direct and diffuse reflection and the sky) and transmitted through the water-air interface is collected by the receiver. Figure 6 shows the various processes contributing to the total upwelling solar radiance (L_t), which includes the surface reflected sky radiance (L_s), surface reflected radiance (L_g) due to sun glint, and water-leaving radiance (L_w). The surface-reflected solar radiance is the direct reflection of sun glint and sky radiation from the water surface. The water-leaving radiance (L_w) is the upwelling radiance emerging from the water column, which is a fraction of solar radiation that penetrates the water. The total upwelling solar radiance is written as

$$L_t = L_w + L_s + L_g \tag{3}$$

1) DATA AND METHODS (RADIOMETRIC MEASUREMENTS)

The RAMSES TriOS hyperspectral radiometers were employed for collecting radiometric data from the above-surface and underwater setups. The above-surface dataset (Fig. 6 (b)) includes the measurements of sky radiance (L_{sky}), downwelling irradiance ($E_d(0^+)$), and total upwelling radiance (L_t) and the below-surface measurements (Fig. 6 (c)) include the downwelling irradiance $E_d(0^-)$ and upwelling radiance $L_u(0^-)$. The in-situ measurements of these properties were collected from Bay of Bengal waters during the daytime from 12 pm to 4 pm (local time, IST). While the direct measurements of L_t are available within the dataset, these data do not provide sufficient insights into the influence of specific parameters like L_w, L_s and L_g on the solar power captured by the receiver. As illustrated in Fig. 6 (a), these radiance (L_w, L_s and L_g) values are influenced by numerous marine environmental factors (including sea surface roughness, wind speed, sun glint intensity, and bubble size/concentrations near the sea surface). Thus, it becomes essential to thoroughly investigate the effects of these parameters on the modelling of the underwater-to-air OWC system's performance. Our in-situ measurements strictly followed the ocean optics protocols and methods presented in Mobley [42], where the radiance sensor was fixed at an angle of 40° from nadir and approximately 135^0 azimuth from the sun (see Fig. 6 (b)). The surface-reflected radiance (L_s) can be calculated based on the L_{sky} values using

$$L_s = \rho L_{sky} \tag{4}$$

Here, ρ represents the Fresnel reflectance coefficient at the air-sea interface, indicating the proportion of light reflected

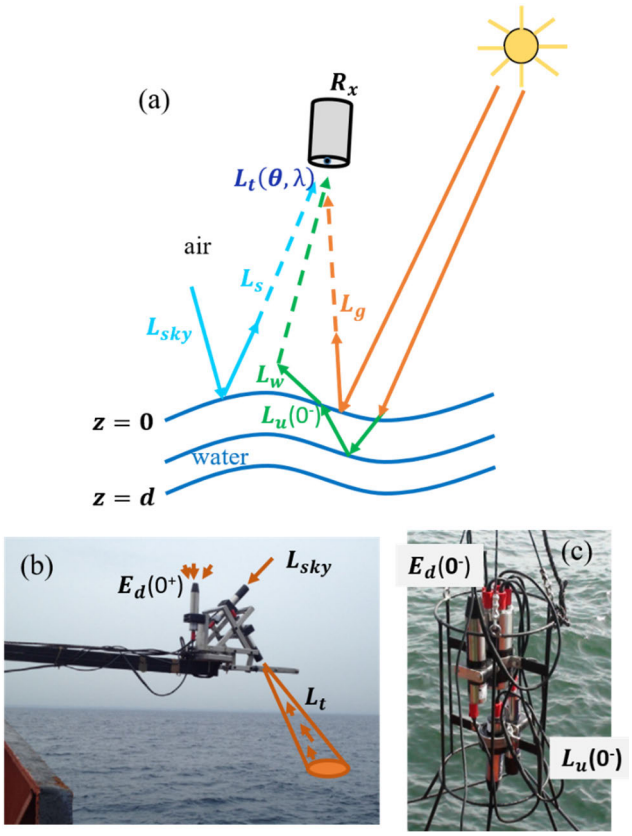


FIGURE 6. (a) Illustration of various processes contributing to the total upwelling solar radiance, (b) Field setup of RAMSES-Trios above-surface radiometers (downwelling irradiance E_d (0^+), sky radiance L_{sky}, and total upwelling radiance L_t) on an extended platform for measurements, (c) Field setup of RAMSES-Trios underwater radiometers installed on an underwater frame for measuring downwelling irradiance E_d (0^-) and upwelling radiance L_u (0^-) below the sea surface.

from the sea surface. Its value typically varies with the factors such as wavelength, wave shape, wind speed, solar zenith angles, and specific viewing geometry [42], [43]. In the case of a flat sea surface and a uniform sky radiance distribution, the value of ρ is 0.02 for a sensor viewing angle (θ) less than 30^0 and increases to 0.03 when θ equals 40^0 [44]. However, the sea surface is often wave-roughened through the combined influence of wind's energy transfer to waves, the inherent turbulence within these waves, and the length of the wind's path over the water. Consequently, this variability underscores that the parameter ρ cannot remain constant and is often greater than the values mentioned earlier. Thus, ρ cannot be a constant value and it is larger than the reported values. Additionally, it is clear that determining this parameter across diverse wind and sea conditions presents a significant scientific challenge. Because of this complex nature, many researchers have considered a fixed value of $\rho = 0.028$ as previously suggested by Mobley [42]. This is typically done under the conditions of a uniform sky radiance distribution, low wind speeds (less than 5 m s^{-1}), viewing angles not exceeding 40^0, and with no influence from sunlight reflection off the sea surface (sun glint). In this study,

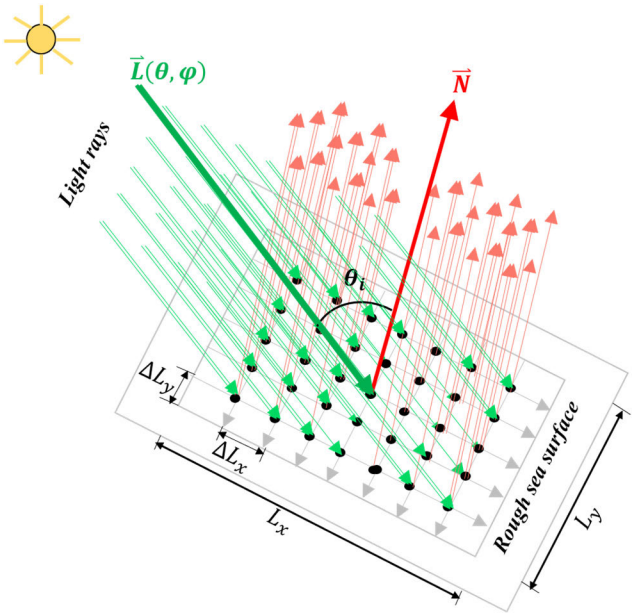


FIGURE 7. Schematic diagram of calculating ρ from the sea surface model.

we determined the ρ value by employing the ECKV method to model the sea surface. We conducted these calculations for a range of solar zenith angles and wind speeds. This approach was necessary because the sea surface roughness varies with changing wind speed and direction, and the intensity of incident sunlight on the sea surface varies with changing solar zenith angles and time intervals. Solar zenith angles can be determined by considering both the time of day and the geographical location and hence vary accordingly. The schematic diagram of calculating the ρ values from the sea surface model is shown in Fig. 7. When parallel sunlight strikes the mesh grid of the sea surface, we calculate the incident angle ($\boldsymbol{\theta_i}$) between the incoming sun rays and the normal vector associated with the mesh grid point on the sea surface. The solar zenith angle (θ) is determined from the time of day, and the azimuthal angle (φ) is randomly generated within the interval of $[0, 2\pi]$. The ρ value is subsequently determined using Fresnel's equations, based on this angle of incidence.

In our study, we considered wind speeds at 3, 5, and 10 m s^{-1}. To accommodate the wavelength scales of capillary and gravity waves, we considered maximum wavelengths (L_x and L_y) of 8, 25, and 100 m for wind speeds of 3, 5 and 10 m s^{-1} respectively, based on a previous study [45]. To maintain the optimal grid resolution (ΔL_x and ΔL_y) of 0.09 cm, we sampled the sea surface at rates of 8192 × 8192, 32768× 32768, and 131072 × 131072 grid points for wind speeds of 3, 5, and 10 m s^{-1} respectively, depending on the values of L_x and L_y. To enhance the accuracy of determining the ρ values, we utilised 1000 different sea surface realisations modelled using the ECKV method for each wind speed in our simulations. Table 1 presents the average reflectance (ρ) value for various solar zenith angles and wind speeds. The calculations revealed that the ρ value is not constant, as previously assumed in earlier studies, but it varies with both the position of the sun and wind speeds. As the wind speed increases, the roughness of the sea surface also increases, causing the incident angles to diverge further and consequently leading to an increase in the average reflectance value. Further, it was observed that the ρ value increases as the solar zenith angle increases.

TABLE 1. The average reflectance (ρ) values for various solar zenith angles and wind speeds.

Solar zenith angles (θ_s)	Wind speed		
	3 m s^{-1}	5 m s^{-1}	10 m s^{-1}
25^0	0.035	0.052	0.104
30^0	0.043	0.063	0.111
45^0	0.081	0.100	0.129
60^0	0.139	0.146	0.149

As shown in Fig. 6 (a), the water-leaving radiances (L_w) were calculated using the underwater radiometer measurements by incorporating sea surface roughness on the upwelling radiance $L_u\left(0^-\right)$ just below the sea surface by using the following expression

$$L_w = \tau L_u\left(0^-\right) \tag{5}$$

where τ is the transmittance coefficient relative to the wave facet and is calculated using Snell's and Fresnel equations. As previously discussed, when calculating the value of τ, we applied the same principles as those used for calculating the ρ values. This implies that τ varies in response to changes in sea surface roughness and wind speed.

2) IMPACT OF BUBBLES ON THE RADIOMETRIC MEASUREMENTS

Wave breaking is a common phenomenon in open ocean waters, predominantly driven by winds. It results in the formation of bubbles on the sea surface, and the depth to which these bubbles penetrate the water column is influenced by both the characteristics of wind-generated waves and the concentrations of water constituents. To examine the impact of bubbles on the radiometric measurements, radiometric measurements were conducted continuously for approximately 3 minutes at a specific location in Bay of Bengal waters. For analytical purposes, the average data from these 3-minute measurements were utilised. These experiments were conducted on a bright sunny day, specifically from 12:00 pm to 4:00 pm (local time, IST), under two different water state conditions: calm water and wave-breaking-induced bubbles, as depicted in Fig. 8 (a) and (b). These measurements were collected at four different times: 12:20 pm, 1:25 pm, 2:45 pm, and 3:50 pm.

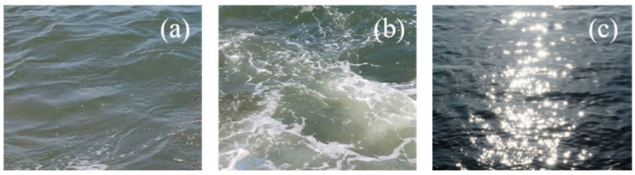


FIGURE 8. Images from field experiments in Bay of Bengal waters on February 15, 2020, between 12 pm and 3 pm, illustrating calm water and bubble cloud formation resulting from wave breaking. (a) calm water; (b) wave-induced bubble clouds; and (c) glint effect.

Figure 9 (a)-(c) displays the spectral sky radiance, downwelling irradiance, and upwelling radiance just below the sea surface at various time intervals in Bay of Bengal waters (all measurements made under clear sky conditions). Using the recorded time data and geographic location, we calculated the respective solar zenith angles (θ_s). These measurements clearly indicate a decreasing trend in spectral radiance and irradiance values as the daytime progresses. This decline in radiance and irradiance values is a direct result of the diminishing intensity of solar light, which decreases notably from 12:20 pm to 4 pm as the sun's position changes throughout the day. In Fig. 9 (c), the solid line corresponds to the upwelling spectral radiance observed under calm water conditions, while the dashed line represents the upwelling radiance spectrum when wave-breaking-induced bubbles formed. These distinct lines provide a clear contrast between the two scenarios, allowing for a comprehensive comparison of radiance spectra between calm and bubble-filled water states. The wave-breaking-induced bubbles typically result in higher radiance values compared to a calm water state because the bubbles scatter and reflect sunlight more effectively. The increased scattering and reflection contribute to higher radiance values.

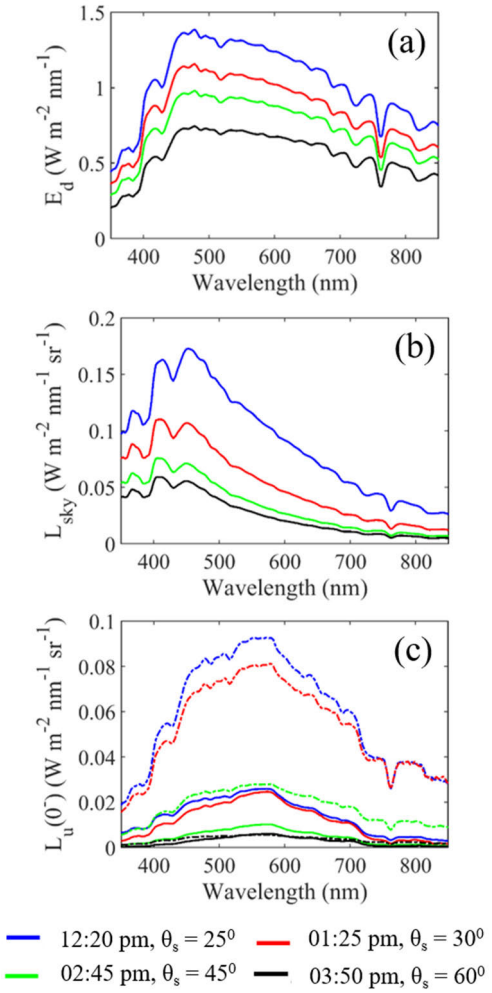


FIGURE 9. In-situ measurements collected at four different daytimes under clear sky conditions: (a) downwelling irradiance (E_d (0^+)), (b) sky radiance (L_{sky}), and (c) upwelling radiance (L_u (0^-)) just below the sea surface. Solid line - calm water radiance, dashed line - wave-breaking-induced bubbles radiance.

3) IMPACT OF SUN GLINT ON RADIOMETRIC MEASUREMENTS

Sun glint refers to the bright, reflective appearance of the sea surface when sunlight is directly reflected off it and towards an observer (such as a satellite or aircraft). This effect occurs due to the specular reflection of sunlight, where light is reflected at the exact angle it strikes the sea surface (as illustrated in Fig. 8 (c)). Even though the radiometric measurements are obtained taking the aforementioned precautions, it is still possible to detect sunlight reflected (sun glint) by the sea surface in the L_t data. Consequently, the surface reflected solar radiance encompasses not just the reflected sky light but also the sunlight reflected (sun glint) from the sea surface. In [44], the sun glint is calculated from the downwelling irradiance using

$$L_g = mE_d\left(0^+\right) \tag{6}$$

The parameter 'm' is a proportionality factor that signifies the magnitude of $E_d\left(0^+\right)$ in the in-situ radiometric data. It ranges between 0 and 0.07, with 'm' equalling 0 indicating the absence of glint conditions, while the higher values of 'm' correspond to the increased sun glint strength. Thus, the total upwelling radiance L_t reaching the sensor is expressed as

$$L_t = \tau L_u\left(0^-\right) + \rho L_{sky} + mE_d\left(0^+\right) \tag{7}$$

Figure 10 shows the variation in L_t as influenced by sun glint across a range of m values.

E. CALCULATING THE SYSTEM'S SIGNAL-TO-NOISE RATIO (SNR) AND BER PERFORMANCE USING AVALANCHE PHOTO DIODE (APD) RECEIVER

In our simulations, we utilised an Avalanche Photo Diode (APD) as a receiver to evaluate the BER performance of an underwater-to-air OWC system. The various electrical noises produced at the UOWC receiver are due to the thermal, shot and dark current noises. In [34], a complete model is presented for analysing the received noise in UOWC system. We used this model to calculate the received electrical noises for the case of APD detector. The APD receiver used

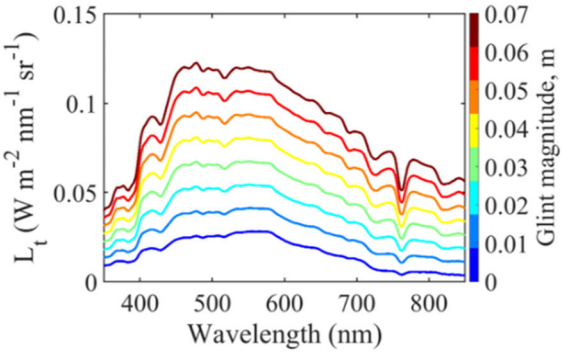


FIGURE 10. Variation in L_t as influenced by sun glint across various values of m.

the approximated zero mean Gaussian model and Poisson model for thermal and shot noises respectively. We reasonably assumed the dark current (I_d) values for APD as 10 nA. The thermal noise variance is given as

$$\sigma^2_{thermal} = \frac{4k_B T_m B_n}{R_{TIA}} \tag{8}$$

where B_n is the APD 's electrical bandwidth expressed in Hz, T_m is the medium temperature, R_{TIA} is the resistance of the trans-impedance amplifier's feedback resistor and k_B is Boltzmann constant.

Solar power reception is determined through the mathematical calculations using solar radiance spectra. This quantity, however, varies significantly based on several crucial parameters, including geographical location, time of day, the elevation or depth of the receiving equipment, and the aperture area's dimensions that are utilised for solar energy collection. The expression for the received solar power ($P_{R-solar}$) at the receiver is given as follows:

$$P_{R-solar} = L_t.T_\theta.A_{Apd}.\Delta\lambda.S \tag{9}$$

where T_θ represents the transmission efficiency of the filter, A_{Apd} is the effective area (m^2) of the receiver, $\Delta\lambda$ denotes the optical filter band width (nm), and S is the solid angle of the optical filter, measured in steradians. The expression for S in terms of the field of view (FOV) of the photo detector can be calculated using Eq. 10 and is given as follows:

$$S = 2\pi\left(1 - \cos\left(\frac{FOV}{2}\right)\right) \tag{10}$$

The variance of shot noise at the receiver due to the reception of both solar ($P_{R-solar}$) and signal power ($P_{R-signal}$) can be expressed as

$$\sigma^2_{shot} = 2qM^2_{APD}\begin{bmatrix} R_{APD}\left(P_{R-signal} + P_{R-solar}\right) \\ +I_d \end{bmatrix} B_n \tag{11}$$

where q is the electron charge, M_{APD} is the APD gain, R_{APD} is the responsivity of APD.

In addition to these two noises, two electrical beat noises ($\sigma^2_{sig-sol}, \sigma^2_{sol-sol}$) are generated at the detector using

$$\sigma^2_{sig-sol} = 4M_{APD}R^2_{APD}P_{R-signal}P_{R-solar}\frac{B_n}{BW_{opt}} \tag{12}$$

$$\sigma^2_{sol-sol} = R_{APD}\left(P_{R-signal}\right)^2\left[2 - \frac{B_n}{BW_{opt}}\right]\frac{B_n}{BW_{opt}} \tag{13}$$

where BW_{opt} is the bandwidth of the optical filter expressed in Hz.

The total variance of noise that occurs during the transmission of optical signal (bit '1') and no optical signal (bit '0') is expressed as

$$\sigma^2_{total,1} = \sigma^2_{shot} + \sigma^2_{thermal} + \sigma^2_{sig-sol} \tag{14}$$

$$\sigma^2_{total,0} = \sigma^2_{thermal} + 2qM^2_{APD}\left[R_{APD}\left(P_{R-solar}\right) + I_d\right]B_n \tag{15}$$

The SNR of the underwater-to-air OWC system is

$$SNR = \frac{\left[M_{APD}R_{APD}P_{R-signal}\left(1 - r_e\right)\right]^2}{\sigma^2_{total,1} + \sigma^2_{total,0}} \tag{16}$$

where r_e is the extinction ratio between optical power associated with bit '0' and '1', and the value is assumed as 0.1.

The performance of the underwater-to-air OWC system using OOK modulation can be expressed as [46]

$$BER = \frac{1}{2}erfc\left(\frac{\sqrt{SNR}}{2\sqrt{2}}\right) \tag{17}$$

where $erfc$ is the complementary error function.

III. RESULTS AND DISCUSSION

To quantify the effect of solar noise on the performance of the underwater-to-air OWC system, numerical results are presented for a range of scenarios, diverse oceanic conditions, varying receiver configurations, and atmospheric conditions. The parametric values for the transmitter and receiver used in our simulations are specified in Table 2. The selection of a green wavelength, specifically 515 nm, offers advantages for an UOWC system due to reduced absorption and enhanced signal transmission at this wavelength. This study employed the MC method using in-situ optical measurement data as the input parameters. Additionally, the HN model was integrated to account for bubble populations during the wave breaking events. In our simulations, the transmitter was situated 50 m below the sea surface and the receiver positioned 2 m above the sea surface. It was assumed that the transmitter is aligned with the receiver. In order to achieve a realistic wireless optical communication through the underwater-to-air system, it is necessary to analyze the impacts of air-sea interface and bubbles in the upper ocean layer.

The sea surfaces were modelled under different wind conditions, specifically at wind speeds of 3, 5 and 10 m s^{-1}. To generate these varied sea surfaces, The ECKV model was employed which accurately replicates the complex dynamics of ocean waves influenced by the varying wind speeds. The

spatial grid resolution for the sea surfaces generated in our simulations is an important consideration in the present study. When making the simplified assumption that the sea surface is perfectly flat, there are no spatial or temporal variations in its characteristics. Consequently, the power received by this flat surface remains constant regardless of the changes in time, wind speed or spatial location. Numerous studies have employed the Cox and Munk sea surface model to analyse the received power in wave surface conditions. However, it is important to note that this model does not provide a comprehensive explanation of the spatial and temporal dynamics of sea surfaces and wave shapes. When the model reveals variations in the received power with varying wind speeds, it is worth emphasising that for a single wind speed scenario, the received power remains constant. This suggests that the Cox and Munk model captures wind-dependent fluctuations in the received power without accounting for the intricate variations in the sea surface characteristics and waveforms that occur under these conditions.

In reality, sea surfaces exhibit a diverse range of wave shapes at different locations, which can encompass crests, troughs, and various random patterns. These complex waveforms introduce dynamic variations in the sea surface over time, making it a dynamic and ever-changing environment. Consequently, the received power from such dynamic sea surfaces can fluctuate considerably due to their temporal and spatial changes.

To investigate the influence of wave shape on the received power, the maximum wavelength (L_x and L_y) was divided into 500 cross-sectional segments for each wind speed. After travelling 50 m through the water, the optical beam passed through the segmented wave cross sections. Observed was a range of fluctuations in the laser beam's intensity at the receiver and this intensity variation was directly related to the specific wave shapes encountered on the sea surface. Consequently, the power received by the detector varied significantly depending on the specific characteristics of the encountered waves, highlighting the substantial impact of wave shape on the received power.

A. NORMALISED RECEIVED POWER

In our analysis, we examined the wind speeds of 3, 5, and 10 m s^{-1} and calculated the received power at the detector plane for each scenario by considering every cross section of the sea surface. The receiver aperture size FOV were considered as 4'' and 180^0 respectively. Our calculations revealed that the presence of waves can significantly impact the received power, potentially resulting in a reduction in the received power and consequently limiting the data transfer capability of the system.

In Fig. 11, we present a histogram depicting the fitting of normalised received power derived from the contributions of 500 distinct cross-sections under varying wind speeds. Our analysis revealed that the received power is not consistent for a specific wind speed, but it falls within a defined range that is influenced by the wind speed. When the wind speed reaches 10 m s^{-1}, the received power is observed to be lower compared to the cases with wind speeds 3 and 5 m s^{-1}. This observation suggests that with the increasing wind speed, the sea surface becomes rougher and leads to a more divergent orientation of the sea surface normal. Consequently, this variation of the sea surface roughness is directly reflected in the received power calculations, which highlight the sensitivity of the system's performance depending on the wind speed and sea surface conditions.

In addition, we investigated the received power variations in the presence of bubbles in the uppermost layers of the

TABLE 2. The Parameters used for MC simulations.

Parameter	Value
Transmitted beam power (P_{Tx})	50 mW
Transmitted beam divergence angle (θ)	1.5 mrad
Transmitted beam width (ω)	1 mm
Transmission beam wavelength (λ)	515 nm
Optical filter bandwidth ($\Delta\lambda$)	15 nm
Optical filter transmissivity (T_f)	1
Gain of APD (M_{APD})	50
TIA resistance (R_{TIA})	50 Ω
Responsivity of APD (R_{APD})	0.8 AW^{-1}
Receiver aperture	4 inches
Receiver FOV	90^0 - 180^0

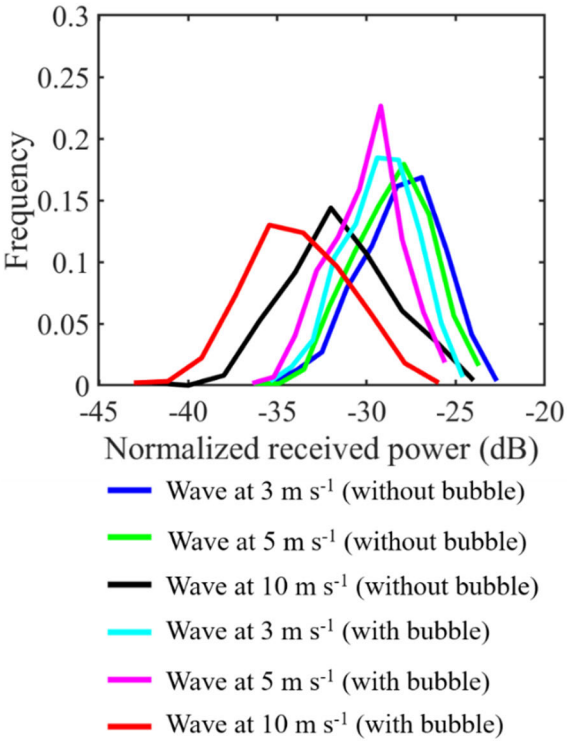


FIGURE 11. The histograms of the normalised received power (aperture: 5'', FOV: 180^0) for 500 sea surface cross-sections at various wind speeds.

sea surface. These bubbles penetrate into the subsurface layer depending on the wind speed. For the MC simulations, we determined the depth to which bubbles penetrated using an exponential decay model that considers the decreasing bubble concentrations with increasing water depth. Our analysis revealed that the wave-breaking-induced bubbles penetrate to water depths of 5, 7, and 10 m for wind speeds of 3, 5, and 10 m s^{-1} respectively.

Figure 11 shows the histograms of the normalised received power (with a 5'' aperture and a FOV 180^0) calculated based on the data collected from 500 different sea surface cross sections and all of which incorporated the bubbles. Our findings clearly demonstrate that the oceanic bubbles have a substantial impact on the normalised received power. When compared to the scenarios without bubbles, the received power significantly decreases under these conditions. Further observations revealed that the oceanic bubbles generated at various wind speeds introduce significant variations in the received power due to changes in the number density of bubbles according to wind speed fluctuations. The bubble number density fluctuation is of utmost importance for UOWC studies, as it directly influences both the scattering and attenuation coefficients. Specifically, at higher wind speeds, these coefficients are increased leading to the increased photon loss owing to a greater number of scattering events. A contributing factor to this phenomenon is that a photon scatters off a bubble at the higher angles when compared to other particulates in the water column. This scattering behaviour causes the optical beam to spread more quickly and widely, ultimately leading to a reduction in the received power intensity at the center of the beam.

B. UNDERWATER-TO-AIR OWC SYSTEM'S BER PERFORMANCE

This section presents a detailed analysis of numerical results to precisely quantify the influence of solar noise on the performance of a typical underwater-to-air OWC link operating in clear oceanic water conditions. We conducted a comprehensive assessment of the system's performance – specifically focusing on BER under various solar noise scenarios. These scenarios were dependent on the factors such as wind speeds, time of day, and bubbles in the upper ocean layer. It is important to note that our analysis assumes the ideal conditions, including perfect beam alignment and precise time synchronisation between the transmitter and the receiver. Data transmission is executed using NRZ-OOK modulation with a transmission speed of 500 Mbps, which allows us to rigorously evaluate the system's performance in realistic operational settings. In the system that employs NRZ-OOK modulation, the APD's bandwidth efficiency is set at 1 bit per second per hertz (1 bit/s/Hz). For this analysis, we have considered an APD receiver and the receiver parameters as provided in Table 2. In our study, we focused on a wavelength of 515 nm. To determine solar power for our analysis, we calculated it based on the solar spectrum specifically at this wavelength under investigation.

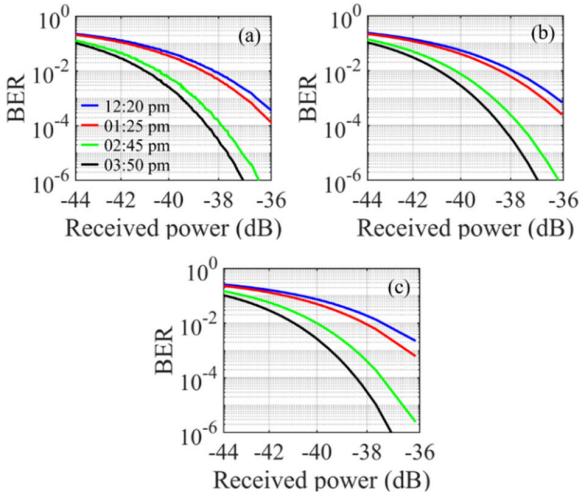


FIGURE 12. BER as a function of the received signal power with an APD receiver and the variations at different times of day and wind speeds (a) 3 m s^{-1}, (b) 5 m s^{-1}, and (c) 10 m s^{-1}.

1) THE IMPACT OF SOLAR NOISE ACROSS DIFFERENT TIMES OF THE DAY

As previously discussed, solar radiation exhibits varying intensities throughout the day. This fluctuation in solar intensity plays a crucial role in influencing the system BER performance in distinct ways at different times of the day. For this analysis, we assumed no bubbles present on the sea surface and performed the BER calculations with the different received signal power for particular wind speed conditions, which determine the shape of the surface waves and hence the solar power variations. Figure 12 presents the BER performance of an underwater-to-air OWC system. It demonstrates how the BER changes in response to the received signal power at different wind speeds and various times of the day. The simulations were executed using a receiver FOV value of 180^0. Our findings indicate that when the Sun passes through the zenith at noon, the BER exhibits poor performance across all wind speed conditions. This deterioration in BER performance is attributed to the increased collection of solar power around midday. As the sea surface wave shapes cause variations in the received signal power, it is imperative to identify the minimum power threshold as necessary to achieve the improved BER performance, aligned with the constraints of the Forward Error Correction (FEC) limit. According to our analysis, we have established specific minimum power requirements to achieve error-free data transmission at various times of the day. These requirements vary depending on the wind speed conditions, ensuring reliable communication in different scenarios.

2) THE IMPACT OF SOLAR NOISE IN THE PRESENCE OF SUN GLINT

The sun glint varies depending on its intensity at different wavelengths, Sun angle and sea surface conditions. In addition to the inherent solar noise from different sources, the sun glint radiation also interacts / interferes with the

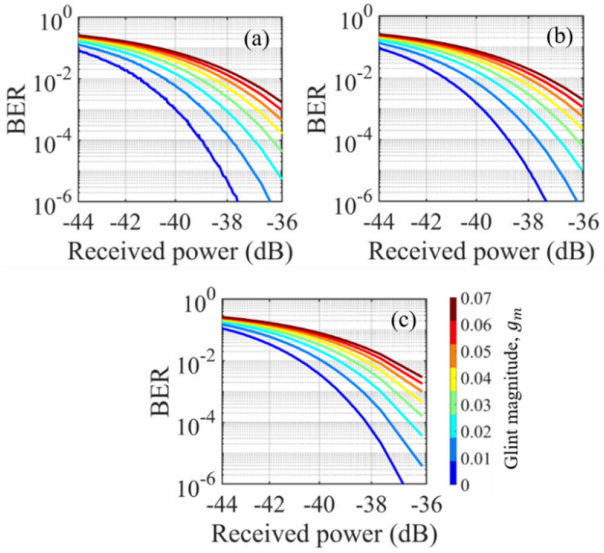


FIGURE 13. BER as a function of the received signal power around midday (noon time) with the varying glint intensities and wind speed conditions (a) 3 m s^{-1}, (b) 5 m s^{-1}, and (c) 10 m s^{-1}.

transmitted optical signal. This phenomenon of sun glint is particularly prominent when the Sun is around midday. In our study, we conducted a detailed analysis of the sun glint effect on the optical signal and system BER performance. For this analysis, the FOV was set at 90^0. Figure 13 shows the BER performance of the OWC system under various sun glint intensities when combined with the other sources of solar noise at wind speeds of 3, 5, and 10 m s^{-1}. The solid blue line on each graph represents the system's BER performance in the absence of sun glint and other solid lines of varying colors depict the BER performance at different sun glint intensities. Our results highlight the substantial impact of sun glint on the system's BER performance. It is noteworthy that if experiments or simulations are conducted on the conditions of maximum possible sun glint, it will become challenging to achieve error-free transmission at the lower received signal power and all wind speeds.

It should be noted that sun glint occurs throughout the day but exhibits higher intensities around midday (at high noon). Consequently, there are fluctuations in the system's BER performance due to the prevailing sun glint conditions. To avoid sun glint during the in-situ experiments in marine environments, one may consider scheduling these measurements during early morning or late afternoon hours when the sun is at a lower angle. Also to adjust the instrument angles and use polarising filters or anti-glare shields to minimise glare from the water's surface.

3) THE IMPACT OF SOLAR NOISE IN THE PRESENCE OF WAVE-INDUCED BUBBLES ON THE SEA SURFACE

Wave breaking is a common occurrence in the marine environments due to the prevailing winds. This oceanographic process generates bubbles in the upper ocean layer and hence it becomes virtually unavoidable to simulate the system's BER performance without accounting the oceanic bubbles.

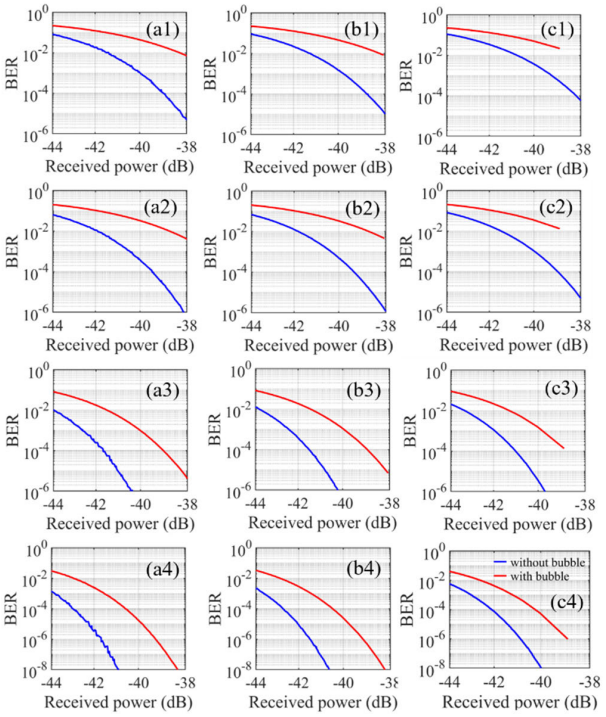


FIGURE 14. BER as a function of the received signal power at different times of the day and wind speeds with and without bubbles. Subplots (a1) through (a4) correspond to a wind speed of 3 m s^{-1}at 12:20 pm, 01:25 pm, 02:45 pm, and 03:50 pm, respectively. Subplots (b1) through (b4) correspond to a wind speed of 5 m s^{-1}, while subplots (c1) through (c4) represent a wind speed of 10 m s^{-1}.

In this study, we made a significant observation from our simulations. The presence of bubbles generated by wave breaking conditions results in distinct variations in the solar power collected at the receiver, in contrast to the scenarios in the absence of bubbles. In our measurements conducted at various times of the day, the wind speed was consistently observed around 5 m s^{-1} when making the measurements of upwelling solar radiance ($L_u\left(0^-\right)$). There were no $L_u\left(0^-\right)$ data collected for instances when bubbles were generated at wind speeds of 3 and 10 m s^{-1}. For these two wind speed conditions, we made the assumption that the solar radiation at the detector was similar to that at a wind speed of 5 m s^{-1}. Nevertheless, the transmittance parameter was taken into account that corresponds to the sea surfaces generated by wind speeds of 3 and 10 m s^{-1}. As a result, the solar noise reaching the detector differed from the condition at 5 m s^{-1} due to the variability of the transmittance parameter associated with the wind-generated sea surfaces. Figure 14 shows the BER performance of the underwater-to-air OWC system for two scenarios: with bubbles and without bubbles at different wind speeds, while considering a receiver FOV of 90^0. According to these simulations, the presence of bubbles in the uppermost ocean layer leads to a reduction in the transmitted power at the detector, and these bubbles have the effect of enhancing the solar radiation reaching the receiver. Consequently, the BER performance deteriorates when bubbles are present at the sea surface as compared to the system's performance in the absence of bubbles.

IV. CONCLUSION

A robust model was presented in this study for evaluating the performance of underwater-to-air OWC system using Monte Carlo technique. Using the simulation and experimental results, we comprehensively assessed the impact of solar noise on the system BER performance by taking into account the various marine environmental factors such as bubbles, wavy surface conditions, and solar noise under different atmospheric and marine conditions. To ensure an accurate simulation of the system's performance, we incorporated the in-situ optical measurements, including IOPs and solar radiance data obtained from Bay of Bengal waters. Additionally, the measurement data were collected from the Southern Ocean to derive and account for the bubble properties using the theoretical HN model. The modelling of wavy surface was achieved through the utilisation of the ECKV model, which incorporates both spatial and temporal dependencies in the generation of sea surfaces. This approach was crucial for our understanding of the optical signal behaviour when propagating through the air-sea interface. This study highlighted the significant influence of both sea wave shapes and bubbles on the received signal power of a laser beam at different wind speed conditions. The findings indicate that the received signal power varied depending on the wind speed and surface wave structure at the air-sea boundary. Additionally, the oceanic bubbles were found to decrease the received signal power as the wind speed increased. Finally, we analysed the system's BER performance under different times of the day, varying sun glint intensities, and different receiver setups. Our findings indicate that the BER performance is significantly deteriorated around midday (when the sun at noon) and improved at other times of the day. Additionally, we observed that the sun glint has a significant impact on the system's BER performance. The higher intensities of sun glint result in reduced BER performance, leading to an increased occurrence of data transmission errors. Additionally, our analysis revealed that the presence of bubbles in the upper ocean layer enhances the reception of solar signal power at the receiver. Consequently, this increase in solar power due to the oceanic bubbles leads to a reduction in the system's BER performance as compared to scenarios in the absence of bubbles. These findings offer essential guidance for system designers as they create and refine underwater-to-air OWC systems by taking into account the various marine and atmospheric conditions encountered in real-world applications.

REFERENCES

[1] A. Morel, B. Gentili, H. Claustre, M. Babin, A. Bricaud, J. Ras, and F. Tièche, "Optical properties of the 'clearest' natural waters," *Limnol. Oceanogr.*, vol. 52, no. 1, pp. 217–229, Jan. 2007, doi: 10.4319/lo.2007.52.1.0217.

[2] K. Nakamura, I. Mizukoshi, and M. Hanawa, "Optical wireless transmission of 405 nm, 145 Gbit/s optical IM/DD-OFDM signals through a 48 m underwater channel," *Opt. Exp.*, vol. 23, no. 2, p. 1558, 2015, doi: 10.1364/oe.23.001558.

[3] D. B. Kilfoyle and A. B. Baggeroer, "The state of the art in underwater acoustic telemetry," *IEEE J. Ocean. Eng.*, vol. 25, no. 1, pp. 4–27, Jan. 2000, doi: 10.1109/48.820733.

[4] Y. R. Zheng, "Channel estimation and phase-correction for robust underwater acoustic communications," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2007, pp. 1–6, doi: 10.1109/MILCOM.2007.4454811.

[5] A. I. Al-Shamma'a, A. Shaw, and S. Saman, "Propagation of electromagnetic waves at MHz frequencies through seawater," *IEEE Trans. Antennas Propag.*, vol. 52, no. 11, pp. 2843–2849, Nov. 2004, doi: 10.1109/TAP.2004.834449.

[6] X. Che, I. Wells, G. Dickers, P. Kear, and X. Gong, "Re-evaluation of RF electromagnetic communication in underwater sensor networks," *IEEE Commun. Mag.*, vol. 48, no. 12, pp. 143–151, Dec. 2010, doi: 10.1109/MCOM.2010.5673085.

[7] H. Kaushal and G. Kaddoum, "Underwater optical wireless communication," *IEEE Access*, vol. 4, pp. 1518–1547, 2016, doi: 10.1109/ACCESS.2016.2552538.

[8] A. S. Fletcher, S. A. Hamilton, and J. D. Moores, "Undersea laser communication with narrow beams," *IEEE Commun. Mag.*, vol. 53, no. 11, pp. 49–55, Nov. 2015, doi: 10.1109/MCOM.2015.7321971.

[9] F. Hanson and S. Radic, "High bandwidth underwater optical communication," *Appl. Opt.*, vol. 47, no. 2, p. 277, 2008, doi: 10.1364/ao.47.000277.

[10] P. Rozmej, A. Karczewska, and E. Infeld, "Superposition solutions to the extended KdV equation for water surface waves," *Nonlinear Dyn.*, vol. 91, no. 2, pp. 1085–1093, Jan. 2018, doi: 10.1007/s11071-017-3931-1.

[11] C. Cox and W. Munk, "Measurement of the roughness of the sea surface from photographs of the Sun's glitter," *J. Opt. Soc. Amer.*, vol. 44, no. 11, p. 838, 1954, doi: 10.1364/josa.44.000838.

[12] T. Elfouhaily, B. Chapron, K. Katsaros, and D. Vandemark, "A unified directional spectrum for long and short wind-driven waves," *J. Geophys. Res., Oceans*, vol. 102, no. C7, pp. 15781–15796, Jul. 1997, doi: 10.1029/97jc00467.

[13] Y. Du, X. Yang, K.-S. Chen, W. Ma, and Z. Li, "An improved spectrum model for sea surface radar backscattering at L-band," *Remote Sens.*, vol. 9, no. 8, p. 776, Jul. 2017, doi: 10.3390/rs9080776.

[14] C. D. Mobley, "Polarized reflectance and transmittance properties of windblown sea surfaces," *Appl. Opt.*, vol. 54, no. 15, p. 4828, 2015, doi: 10.1364/ao.54.004828.

[15] A. Tanaka, H. Sasaki, and J. Ishizaka, "Alternative measuring method for water-leaving radiance using a radiance sensor with a domed cover," *Opt. Exp.*, vol. 14, no. 8, p. 3099, 2006, doi: 10.1364/oe.14.003099.

[16] K. G. Ruddick, K. G. Ruddick, K. Voss, E. Boss, A. Castagna, R. Frouin, A. Gilerson, M. Hieronymi, B. C. Johnson, J. Kuusk, Z. Lee, M. Ondrusek, V. Vabson, and R. Vendt, "A review of protocols for fiducial reference measurements of water-leaving radiance for validation of satellite remote-sensing data over water," *Remote Sens.*, vol. 11, no. 19, p. 2198, Sep. 2019, doi: 10.3390/rs11192198.

[17] D. D'Alimonte, T. Kajiyama, G. Zibordi, and B. Bulgarelli, "Sea-surface reflectance factor: Replicability of computed values," *Opt. Exp.*, vol. 29, no. 16, p. 25217, 2021, doi: 10.1364/oe.424768.

[18] R. Sahoo, S. K. Sahu, and P. Shanmugam, "Estimation of the channel characteristics of a vertically downward optical wireless communication link in realistic oceanic waters," *Opt. Laser Technol.*, vol. 116, pp. 144–154, Aug. 2019, doi: 10.1016/j.optlastec.2019.03.023.

[19] R. Sahoo and P. Shanmugam, "Effect of the complex air–sea interface on a hybrid atmosphere-underwater optical wireless communications system," *Opt. Commun.*, vol. 510, May 2022, Art. no. 127941, doi: 10.1016/j.optcom.2022.127941.

[20] Y. Dong, S. Tang, and X. Zhang, "Effect of random sea surface on downlink underwater wireless optical communications," *IEEE Commun. Lett.*, vol. 17, no. 11, pp. 2164–2167, Nov. 2013, doi: 10.1109/LCOMM.2013.091113.131363.

[21] J. Qin, M. Fu, and B. Zheng, "Analysis of wavy surface effects on the characteristics of wireless optical communication downlinks," *Opt. Commun.*, vol. 507, Mar. 2022, Art. no. 127623, doi: 10.1016/j.optcom.2021.127623.

[22] M. S. Islam and M. F. Younis, "Analyzing visible light communication through air–water interface," *IEEE Access*, vol. 7, pp. 123830–123845, 2019, doi: 10.1109/ACCESS.2019.2938522.

[23] T. Lin, N. Huang, C. Gong, J. Luo, and Z. Xu, "Preliminary characterization of coverage for water-to-air visible light communication through wavy water surface," *IEEE Photon. J.*, vol. 13, no. 1, pp. 1–13, Feb. 2021, doi: 10.1109/JPHOT.2021.3054911.

[24] D. Chen, P. Zhao, L. Tang, and M. Wang, "Modeling and oblique transmission characteristics of an underwater wireless optical communication channel based on ocean depth layering," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 4, no. 3, pp. 424–434, 2024, doi: 10.1364/JOSAA.512023.

[25] C. S. S. Shetty, R. P. Naik, U. S. Acharya, and W.-Y. Chung, "Performance analysis of underwater vertical wireless optical communication system in the presence of weak turbulence, pointing errors and attenuation losses," *Opt. Quantum Electron.*, vol. 55, no. 1, pp. 1–22, Jan. 2023, doi: 10.1007/s11082-022-04283-6.

[26] M. Elamassie and M. Uysal, "Vertical underwater visible light communication links: Channel modeling and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6948–6959, Oct. 2020, doi: 10.1109/TWC.2020.3007343.

[27] Y. Lou, J. Cheng, D. Nie, and G. Qiao, "Performance of vertical underwater wireless optical communications with cascaded layered modeling," *IEEE Trans. Veh. Technol.*, vol. 71, no. 5, pp. 5651–5655, May 2022, doi: 10.1109/TVT.2022.3156388.

[28] B. Rao Angara, P. Shanmugam, and H. Ramachandran, "Influence of sea surface waves and bubbles on the performance of underwater-to-air optical wireless communication system," *Opt. Laser Technol.*, vol. 174, Jul. 2024, Art. no. 110652, doi: 10.1016/j.optlastec.2024.110652.

[29] E. Lamarre and W. K. Melville, "Air entrainment and dissipation in breaking waves," *Nature*, vol. 351, no. 6326, pp. 469–472, Jun. 1991, doi: 10.1038/351469a0.

[30] S. A. Thorpe and P. N. Humphries, "Bubbles and breaking waves," *Nature*, vol. 283, no. 5746, pp. 463–465, Jan. 1980, doi: 10.1038/283463a0.

[31] L. Ma, F. Wang, C. Wang, C. Wang, and J. Tan, "Monte Carlo simulation of spectral reflectance and BRDF of the bubble layer in the upper ocean," *Opt. Exp.*, vol. 23, no. 19, p. 24274, 2015, doi: 10.1364/oe.23.024274.

[32] B. R. Angara, P. Shanmugam, and H. Ramachandran, "Underwater wireless optical communication system channel modelling with oceanic bubbles and water constituents under different wind conditions," *IEEE Photon. J.*, vol. 15, no. 2, pp. 1–11, Apr. 2023, doi: 10.1109/JPHOT.2023.3258500.

[33] W. H. Richey, S. Luminance, and J. E. Tyler, "Light in the sea," *J. Opt. Soc. Amer.*, vol. 53, no. 2, pp. 214–233, 1961.

[34] S. K. Mahapatra and S. K. Varshney, "Impact of orientation-based solar light noise on the performance of underwater optical wireless communication system and noise cancellation," *Results Opt.*, vol. 7, May 2022, Art. no. 100214, doi: 10.1016/j.rio.2022.100214.

[35] T. Hamza, M.-A. Khalighi, S. Bourennane, P. Léon, and J. Opderbecke, "Investigation of solar noise impact on the performance of underwater wireless optical communication links," *Opt. Exp.*, vol. 24, no. 22, p. 25832, 2016, doi: 10.1364/oe.24.025832.

[36] F. Xing and H. Yin, "Performance analysis for underwater cooperative optical wireless communications in the presence of solar radiation noise," in *Proc. IEEE Int. Conf. Signal Process., Commun. Comput. (ICSPCC)*, Sep. 2019, pp. 1–6, doi: 10.1109/ICSPCC46631.2019.8960742.

[37] X. Wang, M. Zhang, H. Zhou, and X. Ren, "Performance analysis and design considerations of the shallow underwater optical wireless communication system with solar noises utilizing a photon tracing-based simulation platform," *Electronics*, vol. 10, no. 5, p. 632, Mar. 2021, doi: 10.3390/electronics10050632.

[38] S. K. Mahapatra and S. K. Varshney, "Performance of the Reed–Solomon-coded underwater optical wireless communication system with orientation-based solar light noise," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 39, no. 7, p. 1236, 2022, doi: 10.1364/josaa.453257.

[39] R. S. Keiffer, J. C. Novarini, and G. V. Norton, "The impact of the background bubble layer on reverberation-derived scattering strengths in the low to moderate frequency range," *J. Acoust. Soc. Amer.*, vol. 97, no. 1, pp. 227–234, Jan. 1995, doi: 10.1121/1.412990.

[40] K. Randolph, H. M. Dierssen, M. Twardowski, A. Cifuentes-Lorenzen, and C. J. Zappa, "Optical measurements of small deeply penetrating bubble populations generated by breaking waves in the Southern Ocean," *J. Geophys. Res., Oceans*, vol. 119, no. 2, pp. 757–776, Feb. 2014, doi: 10.1002/2013jc009227.

[41] X. Zhang, M. Lewis, M. Lee, B. Johnson, and G. Korotaev, "The volume scattering function of natural bubble populations," *Limnol. Oceanogr.*, vol. 47, no. 5, pp. 1273–1282, Sep. 2002, doi: 10.4319/lo.2002.47.5.1273.

[42] C. D. Mobley, "Estimation of the remote-sensing reflectance from above-surface measurements," *Appl. Opt.*, vol. 38, no. 36, p. 7442, 1999, doi: 10.1364/ao.38.007442.

[43] Z. Lee, Y.-H. Ahn, C. Mobley, and R. Arnone, "Removal of surface-reflected light for the measurement of remote-sensing reflectance from an above-surface platform," *Opt. Exp.*, vol. 18, no. 25, p. 26313, 2010, doi: 10.1364/oe.18.026313.

[44] P. J. Dev and P. Shanmugam, "A new theory and its application to remove the effect of surface-reflected light in above-surface radiance data from clear and turbid waters," *J. Quant. Spectrosc. Radiat. Transf.*, vol. 142, pp. 75–92, Jul. 2014, doi: 10.1016/j.jqsrt.2014.03.021.

[45] S. Liu, N. Jia, and C. Guan, "A proposed frequency spectrum for fully developed wind waves by revisiting P-M spectrum," *J. Ocean Univ. China*, vol. 10, no. 4, pp. 331–335, Dec. 2011, doi: 10.1007/s11802-011-1804-7.

[46] G. Giuliano, L. Laycock, D. Rowe, and A. E. Kelly, "Solar rejection in laser based underwater communication systems," *Opt. Exp.*, vol. 25, no. 26, p. 33066, 2017, doi: 10.1364/oe.25.033066.

BHOGESWARA RAO ANGARA received the B.Tech. degree in electronics and communication engineering (ECE) from Andhra University, Andhra Pradesh, India, in 2011, and the M.Tech. degree in control and instrumentation from Delhi Technological University (DTU), Delhi, India, in 2015. He is currently a Ph.D. Scholar with the Department of Ocean Engineering and Electrical Engineering, IIT Madras, Chennai, India. His research interest includes underwater wireless optical communications (UWOC).

PALANISAMY SHANMUGAM received the Ph.D. degree in optical remote sensing from Anna University, Chennai, India, in 2002.

He is currently a Professor with the Department of Ocean Engineering, IIT Madras, Chennai, India, where he is involved in the experimental and theoretical studies of the optical properties of natural water, air–sea interface, and subsea optical processes. He has been a principal investigator of several projects funded by the Government of India. His current research interests include developing models for ocean optical studies and retrieval algorithms for multispectral and hyperspectral remote sensing sensors, with an emphasis on applications to sediment dynamics, algal blooms, climate change, and marine environmental and coastal processes.

HARISHANKAR RAMACHANDRAN received the B.Tech. degree from IIT Bombay, in 1982, and the Ph.D. degree from the University of California at Berkeley, both in electrical engineering.

He was a Research Scientist with the Physics Department, University of California at Los Angeles, for five years before returning to India. He was a Scientist with the Institute for Plasma Research, Gandhinagar, from 1993 to 2001. In November 2001, he joined IIT Madras, where he is currently a Professor with the Department of Electrical Engineering. His research interests include optical link design and planning, nonlinear optics, computational optics, edge plasma physics, and computational plasma physics.

CHAVAPATI GOUSE SANDHANI received the B.Tech. degree in civil engineering from API-IIT Nuzvid, RGUKT University, Nuzvid, Andhra Pradesh, India, in 2015, and the M.Tech. degree in environmental engineering and management from IST, JNTUK University, Kakinada, Andhra Pradesh, in 2018.

He has been a Ph.D. Scholar with the Department of Ocean Engineering, IIT Madras, Chennai, India. His research work involves effect of bubble clouds on the remote sensing reflectance and deriving the water color products with and without bubble clouds.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3412927

Photonic Spin Hall Effect based AQI sensor: Effect of concentration of $PM_{2.5}$

Lokesh Ahlawat, Kamal Kishor, Monu Nath Baitha and Ravindra Kumar Sinha

***Abstract*—The current study proposes an Air Quality Index (AQI) sensor based on the Photonic Spin Hall Effect (PSHE). It senses the changes in AQI caused by the variations in the concentration of particulate matter ($PM_{2.5}$). The layered structure utilizes the Surface Plasmon Resonance (SPR) mechanism for PSHE shifting and enhancement. The present study reports an enhanced PSHE shift of 1888 μm. The proposed AQI sensor possesses a sensitivity of 42.0349°/RIU and a Figure of Merit (FOM) of 4.2×10^6/RIU. The PSHE shift ($\delta_{r\pm}^{H}$) also shows a switching response upon change in AQI range which may serve as an indicator of high pollution level.**

***Index Terms*— Refractive Index Sensor, Photonic Spin Hall Effect, Surface Plasmon Resonance, $PM_{2.5}$, Lorentz Lorenz Equation.**

I. Introduction

A family of Hall effects (Quantum Hall Effect, Anomalous Hall Effect, Spin Hall Effect) have been discovered with some winning Nobel Prizes (Integral QHE in 1985, Fractional QHE 1998). Photonic Spin Hall Effect (PSHE) is the most recent member of the Hall Effect family[1]. The PSHE manifest itself as a spin-dependent splitting of linearly polarized light into Right Circularly Polarized (RCP) and Left Circularly Polarized (LCP) components[2]. This splitting occurs due to spin-orbit coupling upon light-matter interaction[3].

Photons carry both Spin Angular Momentum (SAM)[3] and Orbital Angular Momentum (OAM)[4]. When a linearly polarized light is incident on an optical interface SAM is not conserved perpendicular to the incident plane. Therefore, RCP and LCP components of incident light splits up transversely to conserve the total angular momentum of light[5]. The splitting of RCP and LCP perpendicular to incident plane generates External Orbital Angular Momentum (EOAM). The EOAM leads to PSHE.

The PSHE is also explained in terms of the Rytov-Vladimirskii-Berry (RVB) Phase and the Panchratnam-Berry (PB) Phase[6]. When a light beam interacts with an optical interface one of two processes takes place either the change in direction of wavevectors upon reflection/refraction or a change in polarization of incident light. The variation in the wavevector's direction of incident light generates the RVB phase whereas the change in polarization of incident light is explained based on the PB phase. Polarization and trajectory ($\vec{k}$) of light are coupled by the transverse nature of light[7]. If wavevectors are represented as radii of a sphere in the momentum space (k-space) then the electric field ($\vec{E}$) is tangent to this sphere[8]. The change in the direction of the wavevector upon reflection/transmission leads to parallel transport of the electric field vector over the sphere in k-space. This parallel transport causes no local rotation to the electric field but adds a phase to it. The phase generated between the incident and reflected electric field vector due to parallel transport is referred to as the RVB phase (ϕ_B=-σ.θ)[9]. This geometric Berry phase is opposite for RCP (σ=+1) and LCP (σ=-1) because of their eigenvalues and hence causes them to split.

The PSHE is the flag bearer of spin photonics and is sometimes confused with the Imbert-Fedorov (IF) Shift[10] as both originated from Spin-Orbit Interaction (SOI) of light and are perpendicular to the incident plane. The IF shift is an overall centroid shift whereas the PSHE is the splitting of linearly polarized light into RCP and LCP components[5].

The PSHE shift being sensitive to optical parameters finds applications in precision metrology[5], edge detection[11], optical switches[12], sensors[2], mathematical operations[11], nanophotonics and much more. Along with unique applications come obstacles. The PSHE is a nanoscale phenomenon and difficult to observe directly. This limitation was addressed by enhancing the PSHE through various techniques like evaluation of the PSHE shift at Brewster angle[13] and resonance angle[14] which gave rise to Giant PSHE (G-PSHE)[15]. The PSHE has been observed in hyperbolic metamaterials[16], metasurfaces[17], photonic crystals[18], layered structures[19], zero index metamaterials[20], Moiré surfaces[21], [22]. Seeing the universality, wide-scale applicability and integrability makes PSHE an interesting area of exploration. PSHE being sensitive to optical parameters makes it an excellent choice for sensing. Sensors based on light-matter interaction find unique applications in temperature sensing[23], biosensing[24], gas sensing[25] etc. The PSHE sensing has been utilized to distinguish cancer cells from normal cells as well as detect their concentration[26], gas sensing[27], number of graphene layer

Lokesh Ahlawat, Dr.Kamal Kishor and Prof.Ravindra Kumar Sinha are at TIFAC-Centre for Relevance and Excellence in Fiber Optics and Optical Communication, Department of Applied Physics, Delhi Technological University, Delhi, India (e-mail:dr_rk_sinha@yahoo.com).

Dr.Monu Nath Baitha is with School of Mechanical Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:50:46 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3412927

detection[28], haemoglobin sensing[2] and protein sensing[29].

In this paper, variation of concentration of Particulate Matter is studied based on Kretchmann Configuration[2]. Particulate matter i.e., $PM_{2.5}$ and PM_{10} are a type of aerosol (solid or liquid particles in air) having diameters less than 2.5 µm and 10 µm respectively. $PM_{2.5}$ can penetrate deep into the respiratory tract reaching the lungs and entering the bloodstream. Continuous monitoring of $PM_{2.5}$ is required to minimize the exposure of $PM_{2.5}$ to the human population. The AQI[30] serves as an indicator for measuring the concentration of particulate matter in air. The AQI sensors work on the principle of scattering of light by particulate matter. The scattering of light technique is size-dependent and works well for particle sizes greater than 500 nm[31]. Another technique used for characterizing particles is based on mass. In the present article theoretical AQI sensor is proposed. It works on the principle of reflection of light from a layered structure and is independent of the size of the particulate matter. The proposed layered structure utilizes the Krestchmann configuration[2], [12] for the generation of SPR to detect different concentrations of $PM_{2.5}$ (i.e., AQI). The occurrence of PSHE shift at a unique resonance angle for different AQIs gives the proposed design a high sensitivity of 42.0349º/RIU and a high FOM of 4.2×10^6/RIU. The PSHE shift reaches a value of 1888 µm for AQI 340. The proposed sensor is also compared different plasmonic sensors in Table.3 and to different PSHE-based sensors in Table.4.

II. Design and Modelling

To observe PSHE in the proposed structure, a plane-polarized Gaussian beam having a wavelength of 632.8 nm is considered. The prism in the Krestchmann configuration, depicted in Fig. 1, is made of Boron Nitride (BN) with a refractive index of 2.1. Beneath the prism, is a layer of gold with a thickness of 48.5 nm and a complex refractive index of 0.2166 + i3.2322.

At the interface of the prism-Au layer (i.e., $x = 0$), we consider a global Cartesian coordinate frame (x, y, z) attached to the structure. The incident (x_i, y, z_i) and reflected (x_r, y, z_r) beams are considered to have a Laboratory frame coordinate system affixed to them. z_i and z_r signify the direction of propagation of the incident and reflected light respectively. The y-axis points in the same direction for the incident and reflected light beams as shown in Fig.1.

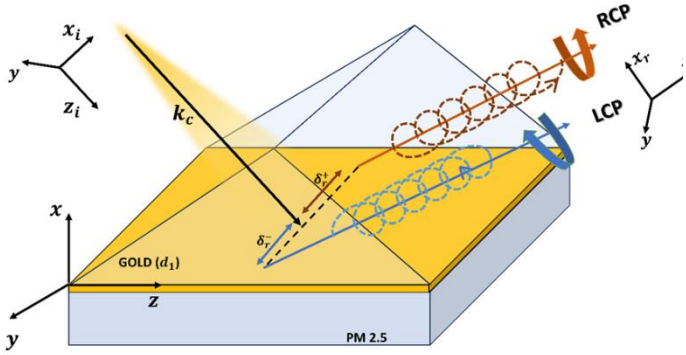


Fig. 1. **The transverse PSHE shift ($\pm\delta_r^+$) of RCP and LCP components of incident light in the Kretchmann configuration comprising of a prism, gold layer (with thickness d_1) and $PM_{2.5}$ layer as the substrate. k_c is the central wavevector of the incident Gaussian beam.**

III. Theory

A. Surface Plasmon Resonance

The leakage of light into the classically forbidden region i.e., from a higher index medium into a lower index medium is referred to as evanescent wave. When light is incident at an angle greater than the critical angle, it leads to the generation of an evanescent wave. This evanescent wave decays exponentially in the axial direction with respect to the interface of the two mediums[32]. Therefore, it can be localized up to a few hundred nanometres, which is the skin depth of the wave. This localized field generates surface selective excitations and is ideally suited for sensing applications. This fact has been utilized to study the variation of the concentration of $PM_{2.5}$ using the proposed layered structure shown in Fig.1.

The generation of evanescent waves, in the case of a dielectric-metal interface, extends into the phenomenon of SPR. The wave travelling along the metal-dielectric interface is known as the Surface Plasmon Polariton (SPP)[33], [34] and the evanescent field decays exponentially normal to the interface.

B. Photonic Spin Hall Effect

A linearly polarized electromagnetic wave is a superposition of RCP and LCP light. Therefore, the electric field vector associated with linearly polarized light can be decomposed into RCP and LCP components[2]. It can be written in the circular basis in the Ket form as:

$$\left|E_H^i\right\rangle = \frac{1}{\sqrt{2}}\left(E_{|+\rangle} + E_{|-\rangle}\right) \quad (1)$$

$$\left|E_V^i\right\rangle = \frac{i}{\sqrt{2}}\left(E_{|+\rangle} - E_{|-\rangle}\right) \quad (2)$$

where H stands for horizontally-polarized (in-plane polarization) which is same as p-polarized and V stands for vertical-polarized (out-of-plane polarization) which is same as light. The RCP and LCP Jones vectors are $E_{|+\rangle}$ and $E_{|-\rangle}$ respectively. The incident Gaussian beam is given as

$$\tilde{E}_{i|\pm\rangle} = \frac{w_0}{2\pi}\left(E_{|x\rangle} + i\sigma E_{|y\rangle}\right).e^{\frac{-w_0^2\left(k_{ix}^2+k_{iy}^2\right)}{4}}.e^{i(k_{iz}.z-\omega t)} \quad (3)$$

Upon interacting with an optical interface splits up transverse to the incident plane, where beam waist is represented by w_0. The eigenvalues of RCP and LCP are σ=±1 respectively. The x and y component of the electric field are denoted by $E_{|x\rangle}$ and $E_{|y\rangle}$ respectively. The x, y and z direction wavevectors are k_{ix}, k_{iy} and k_{iz} respectively with $k_{iz} \gg k_{ix}, k_{iy}$. The perpendicular shift in reflected light is given as:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:50:46 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3412927

$$\langle \hat{y} \rangle = \frac{\langle \tilde{E}_H^r \left| \frac{i\partial}{\partial k_{ry}} \right| \tilde{E}_H^r \rangle}{\langle \tilde{E}_H^r | \tilde{E}_H^r \rangle} \quad (4)$$

where the electric field of the reflected p-polarized light is given by $\tilde{E}_H^r$. Also, $\tilde{E}_H^r$ is related to the incident Gaussian beam vector by the modified reflected Jones matrix as: -

The electric field profile associated with reflected Gaussian beam is represented as: -

$$\tilde{E}_H^r = \frac{r_p}{\sqrt{2}} \left[e^{(-ik_{ry}\delta_{r+}^H)} E_{|+\rangle} + e^{(ik_{ry}\delta_{r+}^H)} E_{|-\rangle} \right] \quad (6)$$

where,

$$\delta_{r\pm}^H = \mp \left(\frac{1}{k_0}\right) \left[1 + \frac{|r_s|}{|r_p|} cos(\varphi_s - \varphi_p) \right] cot\, \theta_i \quad (7)$$

The transverse shift of reflected p-polarized light into RCP and LCP components respectively is $\delta_{r\pm}^H$. The incident wavevector of light is k_0. The Fresnel reflection coefficients for s and p-polarized light are r_s and r_p respectively. The phase attained by the s and p-polarized light after reflection from the optical interface is φ_s and φ_p respectively. θ_i is the angle of incidence.

C. Lorentz Lorenz Equation and AQI sensing

The theoretical modelling of the proposed AQI sensor is based on the Lorentz-Lorenz equation. This equation relates the macroscopic properties of the medium i.e., refractive index and the concentration with the microscopic properties like the polarizability of the scattering particles present in the system. It serves as a bridge between the Maxwell equations and the theory of atomicity[35]. The interaction of the electromagnetic field with the dielectric medium ($PM_{2.5}$ layer in the present studies) leads to the modification of the applied field due to the polarization of the medium. The Lorentz-Lorenz equation is given as: -

$$\frac{4\pi N\alpha}{3} = \frac{n^2 - 1}{n^2 + 2} \quad (8)$$

where, N is the concentration of the particles in the dielectric medium ($PM_{2.5}$ particles for the present study), α is the polarizability and n is the refractive index of the medium.

The concentration of $PM_{2.5}$ particles and the AQI indices have been obtained from[36]. The value of α is taken as 3.75×10^{-9} m^3 (assumed to be constant throughout the range). The concentration values for $PM_{2.5}$ have been plugged into eqn. (8) for calculating the refractive index of the medium. The value of the refractive index of $PM_{2.5}$ varies between 1.3-1.8[37] and it has been taken as 1.8 for the calculation of α. Table.1 reflects the different ranges of AQI indices along with their risk assessment.

	AQI Colour Index	Concern	AQI	Air Quality Description
	Green	Good	0 to 50	Satisfactory air quality.
	Yellow	Moderate	51 to 100	Risk to people sensitive to air pollution.
	Orange	Unhealthy for Sensitive Groups	101 to 150	Risk to people sensitive to air pollution
	Red	Unhealthy	151 to 200	The sensitive group may experience serious health effects
	Purple	Very Unhealthy	201 to 300	Everyone is affected by air pollution level
	Maroon	Hazardous	301 and higher	Emergency situation

Table1. AQI colour index along with the level of risk proposed to different groups (adapted from[38])

IV. Result and Discussion

The incident light is considered to be a plane-polarized Gaussian beam, as seen in Fig.1, interacts with the layered structure. The variation in the concentration of $PM_{2.5}$ alters the refractive index of the substrate. These changes in $PM_{2.5}$ concentration are represented in terms of the AQI index as shown in Table.1. As seen from eqn. (8), an increase in the concentration of particles has a direct impact on the effective refractive index of the structure. Using eqn. (7), PSHE has been calculated for the reflected light in the layered structure for AQI classification. The generation of SPR in this layered structure plays a pivotal role in PSHE enhancement and in obtaining an excellent shift in the resonance angle.

At resonance angle, maximum absorption of incident light takes place in the proposed layered structure due to the generation of SPP and evanescent wave. This exponentially decaying evanescent wave is utilized for concentration detection of $PM_{2.5}$ near the gold layer. An increase in absorption corresponds to a decrease in the Fresnel reflection coefficient, r_p. As seen from eqn. (7), reduction in the r_p enhances the PSHE shift for p-polarized light. The Transfer Matrix Method (TMM)[15] has been utilized for the calculation of $|r_s|$ and $|r_p|$ for the proposed layered structure (Fig.1) in MATLAB.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:50:46 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3412927

A. SPR in the proposed structure

The proposed layered structure is depicted in 2D in Fig.2a where the uppermost cyan coloured layer depicts the Boron Nitride prism, the yellow middle layer is the metallic layer of gold and the bottom-most(grey coloured) layer i.e., the substrate has been taken as $PM_{2.5}$. Fig.2b shows the electric field profile at the Au- $PM_{2.5}$ (AQI 50) interface simulated in COMSOL. It illustrates the decaying evanescent wave normal to the Au-$PM_{2.5}$ interface. Here, the simulations depicted are for AQI 50 but similar results will be obtained for all other values of AQI owing to the same phenomenon as discussed above.

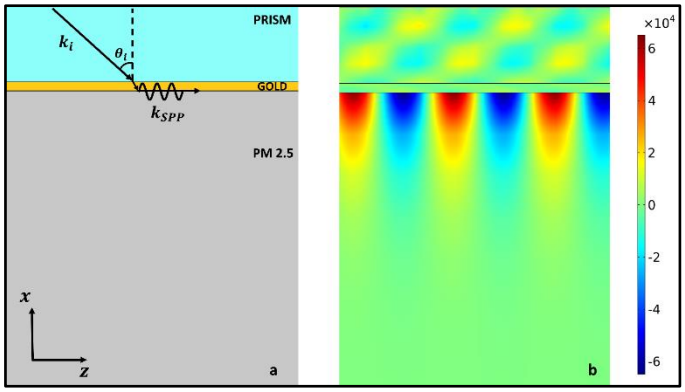


Fig. 2. **SPP at Au-$PM_{2.5}$ (AQI 50) interface for p-polarized light incident at 30.7259º**

The electric field profile corresponding to the decaying field intensity (Evanescent wave) of Fig.2b is shown in the graphical form (blue curve) in Fig.3. The exponentially decaying evanescent wave has a skin depth of 290.1 nm. The sinusoidally propagating electric field (SPP) along the Au- $PM_{2.5}$ interface (green curve) has a wavelength of 589.5 nm, which is less than the incident wavelength (632.8 nm) as expected.

SPR takes place when the horizontal component of incident light wavevector ($n_{Prism}k_0 \cos\theta_i$) matches SPP wavevector (k_{SPP}). Therefore, the Krestchmann configuration has been utilized in the present work which matches k_{SPP} ($(1.0653 \times 10^7\ m^{-1})$ for AQI 50 and horizontal the component of incident p-polarized light at resonance angle of 30.7251º.

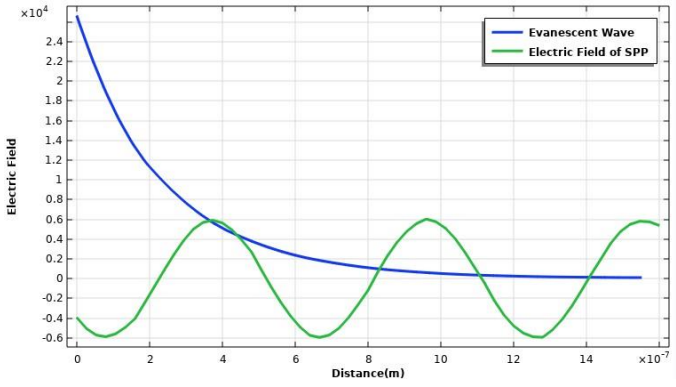


Fig. 3. **Evanescent wave and SPP for AQI 50 in the substrate region**

As shown in Fig.4, there is a sudden dip in the Fresnel reflection coefficient (r_p) because at resonance the absorption light incident on the structure is the largest. When the angle of incidence (θ_i) is greater than the critical angle (θ_C), it leads to the generation of the evanescent wave which penetrates the $PM_{2.5}$ substrates. This evanescent wave is concentrated at the Au- $PM_{2.5}$ interface at the resonance angle (θ_R). The evanescent wave interacts with the $PM_{2.5}$ substrate below the gold surface making the resonance angle sensitive to changes near the gold layer. Changes in the concentration of $PM_{2.5}$ in air lead to changes in the AQI and hence in the proposed structure's effective refractive index.

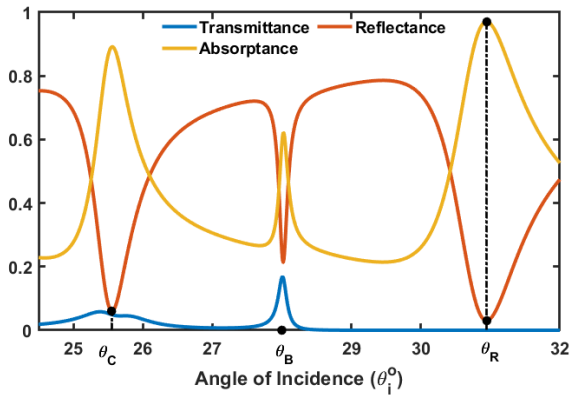


Fig. 4. **Brewster angle (θ_B), critical angle (θ_C) and resonance angle (θ_R) of proposed layered structure for AQI 50**

B. PSHE in the proposed structure

As depicted in Fig.5, the PSHE enhancement has been investigated in the green AQI (0-50) range. The shift in the reflected RCP takes place in the positive y-axis for the whole green range. This shift ($\delta_{r\pm}^H$) is different for different AQIs and occurs at different incident resonance angles owing to SPR. The maximum enhancement ($\delta_{r\pm}^H$= 251.66 µm) takes place for AQI 50 at θ_i= 30.7251º.

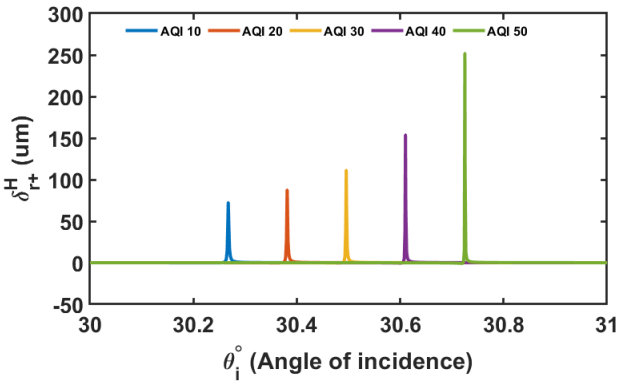


Fig. 5. **Shift in PSHE and resonance angle with change in AQI (10-50)**

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:50:46 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3412927

As seen from Fig.6, the PSHE shift has been investigated for the yellow AQI (60-100) range. The shift in the reflected RCP takes place in the negative y-axis for the whole yellow range. The PSHE shows a switching action in going from green AQI range to yellow AQI range. The PSHE shift ($\delta_{r\pm}^{H}$) is different for different AQIs and occurs at different incident resonance angles owing to SPR. The maximum enhancement ($\delta_{r\pm}^{H}$= 1314 μm) takes place for AQI 60 at θ_i=30.9320º.

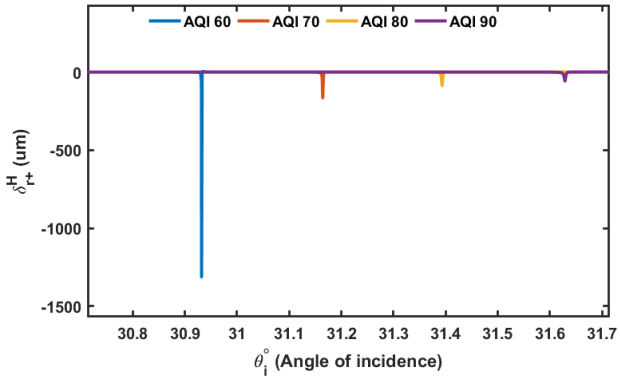


Fig. 6. **Shift in PSHE and resonance angle with change in AQI (60-90)**

As shown in Fig.7, the PSHE shift for the orange (110-150) AQI range has been evaluated. The shift in the reflected RCP takes place in the negative y-axis for the whole orange AQI range. Again, switching of δ_{r+}^{H} from positive y-axis in yellow AQI range to negative y-axis in orange AQI range has taken place. The shift in δ_{r+}^{H} is different for different AQIs and occurs at different incident resonance angles owing to SPR. The maximum enhancement (δ_{r+}^{H}= 36.59 μm) takes place for AQI 110 at θ_i=32.0501º.

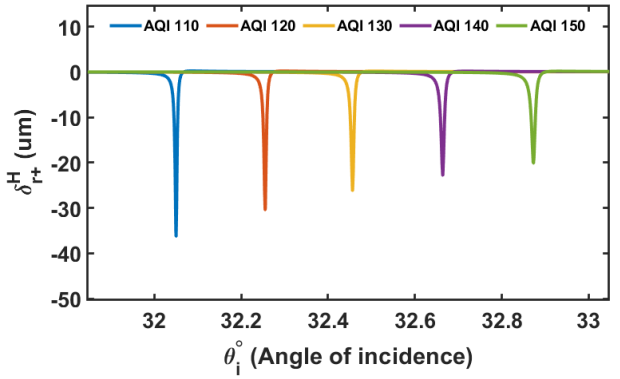


Fig. 7. **Shift in PSHE and resonance angle with change in AQI (110-150)**

The shift in PSHE, as shown in Fig.8, has been investigated for the red AQI (160-200) range. The shift in the reflected RCP takes place in the negative y-axis for the whole red range. This shift in $\delta_{r\pm}^{H}$ is different for different AQIs and occurs at different incident resonance angles owing to SPR. The maximum enhancement ($\delta_{r\pm}^{H}$= 13.38 μm) takes place for AQI 160 at θ_i=33.7794º.

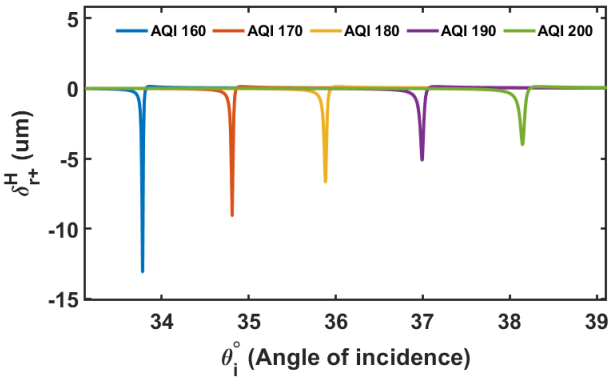


Fig. 8. **Shift in PSHE and resonance angle with change in AQI (160-200)**

As observed in Fig.9, the PSHE shift has been investigated for the purple range of AQI (210-300). The shift in the reflected RCP takes place in the negative y-axis for the whole green range. This shift in $\delta_{r\pm}^{H}$ is different for different AQIs and occurs at different incident resonance angles owing to SPR. The maximum enhancement ($\delta_{r\pm}^{H}$=3.86 μm) takes place for AQI 210 at θ_i=38.7055º.

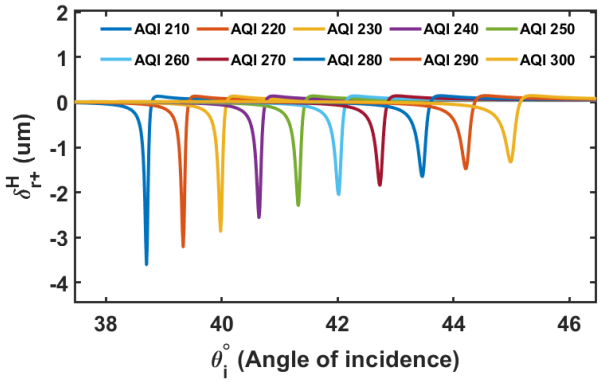


Fig. 9. **Shift in PSHE and resonance angle with change in AQI (210-300)**

As r in Fig.10, the shift in PSHE has been investigated for the maroon AQI (300+) range. The shift in the reflected RCP takes place in the positive y-axis for AQI 310-330 and in the negative y-axis from AQI 340-370. This switching in $\delta_{r\pm}^{H}$ takes place owing to modifications made in the proposed structure. The BN prism is replaced by InAs prism (n_{Prism}=3.966) and gold layer thickness is reduced from 48.5 nm to 46 nm. The modification has been made to observe the significant PSHE enhancement. After these modifications, resonance takes place at a smaller angle of incidence which is caused by increase in the refractive index of prism. The maximum enhancement ($\delta_{r\pm}^{H}$=1888 μm) takes place for AQI 340 at θ_i=23.1705º.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:50:46 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3412927

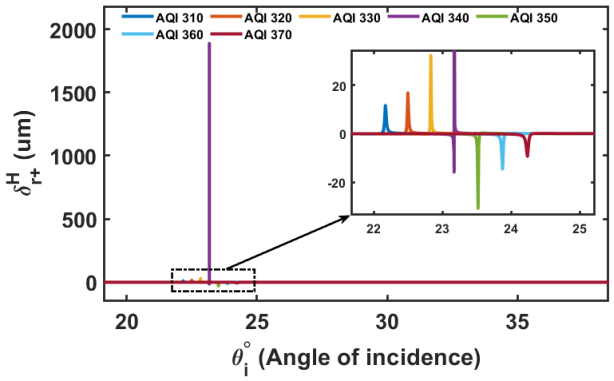


Fig. 10. **Shift in PSHE and resonance angle AQI (300+)**

AQI	RI	FWHM	FOM (RIU^{-1})	PSHE (μm)
50	1.0154	0.0007º	6×10^{4}	251.6
60	1.0209	0.0001º	4.2×10^{6}	1314
110	1.0505	0.0052º	8×10^{3}	36.59
160	1.0951	0.0156º	2.7×10^{3}	13.38
210	1.2138	0.0652º	6.4×10^{2}	3.85
290	1.3322	0.1757º	2.4×10^{2}	1.69
300	1.3477	0.1973º	2.1×10^{2}	1.53
340	1.4096	0.0001º	2.1×10^{5}	1888

Table.2 PSHE based sensors and their PSHE enhancement and FOM

C. Sensitivity and Figure of Merit

Sensitivity (S)[39] for the present work is characterized by the ratio of variation in the resonance angle ($\Delta\theta_R$) corresponding to change (Δn) in the refractive index of $PM_{2.5}$ layer beneath the gold layer. The proposed sensor's sensitivity performance is evaluated using eqn. (9).

$$S = \frac{\Delta\theta_R}{\Delta n} \tag{9}$$

FOM characterizes the performance of the sensor relative to its alternatives and is calculated using eqn. (10).

$$FOM = \frac{S}{FWHM} \tag{10}$$

Here, FWHM stands for Full Width at Half Maximum.

Sensitivity of the proposed sensor is calculated from the slope of Fig.11 and is found to be 42.0349º/RIU. Table.2 summarizes the findings of the presented work. In Table.2, refractive indices for different AQIs and their PSHE enhancement and the performance of the sensor in terms of FOM is reported. AQI 60 and AQI 340 are found to have the same FWHM (i.e. 0.0001º) but different PSHE shift. The PSHE shift reaches its highest value of 1888 μm for AQI 340. The highest value of FOM (4.2×10^{6} RIU^{-1}) is found for AQI 60.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:50:46 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3412927

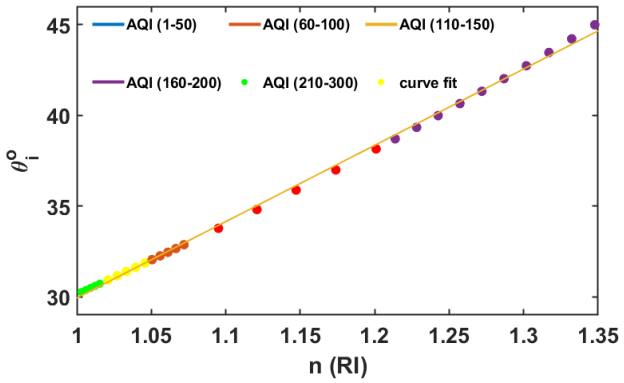


Fig. 11. **Variation of resonance angle (coloured circular dots) for different AQIs and linear curve fit (blue solid line)**

Table.3 compares the performance of the proposed AQI sensor with other plasmonic sensors in terms of sensitivity and FOM. The proposed sensor possesses the highest FOM (4.2×10^6 /RIU) as compared to other plasmonic sensors.

In Table.4 the proposed PSHE based AQI sensor is compared to different PSHE based sensors. The transverse displacement of the RCP of the present model is 12 times more than its contemporary. Hence, qualifying to the category of Giant Photonic Spin Hall Effect (G-PSHE).

Ref.	Structure	λ	S	FOM(RIU^{-1})
[40]	Tunable GST-based Perfect Absorber	1728 nm	900 nm/RIU	15
[41]	Metamaterial based on Plasmonically induced transparency (PIT) effect	200 µm	91,114 nm/RIU	—
[42]	Optical Nano Biosensor-Based Plasmonic Metasurface	1.126 µm	315 nm/RIU	398.8
[43]	Plasmonic Metasurface based on plasmon-induced transparency	1728 nm	497.8	480
[44]	Metamaterial H-shaped dielectric antenna	0.5–2 mm	9 µm/K	—
	Kretchmann Configuration (Present work)	638.2 nm	42.0349°/RIU	4.2×10^6

Table.3 Performance of different plasmonic sensors in terms of Sensitivity and FOM for different wavelengths.

Structure	Material	PSHE (δ_{r+}^H)	RI sensing range	FOM (RIU^{-1})
Modified Otto[27]	Prism|Organic|Graphene|Gas	1882 µm	1.000-1.001	2.3×10^2
1D-PhCR[18]	$(A,B)^{10}\|D\|(A,B)^{10}$ A, D=SiO_2, B=Si_3N_4	35.85 µm	1.460-1.461	6.7×10^9
SPR[2]	Prism|Au|Analyte|Air	150 µm	1.3442–1.3602	3.2×10^3
SPR (Present work)	Prism|Au|$PM_{2.5}$	1888 µm	1.0-1.8	4.2×10^6

Table.4 Performance of various PSHE based sensors in terms of Sensitivity and PSHE Enhancement for a wavelength of 632.8 nm

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:50:46 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3412927

V. Conclusion

In summary, a PSHE based AQI sensor is proposed as a new class of AQI sensors. The proposed sensor utilizes Kretchmann configuration for the generation of SPR. The resonance occurs at unique angle of incidence for all AQIs. Unique resonance angle for each AQI gives our model a sensitivity of 42.0349º/RIU. An enhanced PSHE shift of 1888 μm is reported which is 12 times more than its previously reported results. The presented design possesses FOM of 4.2×10^6 /RIU which is several orders larger than other plasmonic sensors. The PSHE shift shows a switching action upon changing green AQI range to yellow AQI range and a similar switching is seen upon changing yellow AQI range to orange AQI range. This switching behaviour serves as an indicator of increase in pollution levels.

VI. Acknowledgements

Authors gratefully acknowledge the initiatives and support from the TIFAC-Centre of Relevance and Excellence in Fiber Optics and Optical Communication at Delhi College of Engineering, now DTU, Delhi, under the Mission REACH program of Technology Vision-2020, Government of India. One of the authors, Mr. Lokesh Ahlawat, would like to acknowledge (i) Delhi Technological University, Delhi, for financial support in the Research Fellowship, (ii) Ms. Varnam Sherawat of Delhi Technological University, Delhi for critical, analytical and motivating discussions all along this research work.

References

[1] M. Onoda, S. Murakami, and N. Nagaosa, “Hall effect of light,” *Phys. Rev. Lett.*, vol. 93, no. 8, Aug. 2004, doi: 10.1103/PhysRevLett.93.083901.

[2] L. Ahlawat, K. Kishor, and R. K. Sinha, “Photonic spin Hall effect-based ultra-sensitive refractive index sensor for haemoglobin sensing applications,” *Opt. Laser Technol.*, vol. 170, no. June 2023, p. 110183, 2024, doi: 10.1016/j.optlastec.2023.110183.

[3] K. Y. Bliokh, F. J. Rodríguez-Fortuño, F. Nori, and A. V. Zayats, “Spin–orbit interactions of light,” *Nat. Photonics*, vol. 9, no. 12, pp. 796–808, Dec. 2015, doi: 10.1038/nphoton.2015.201.

[4] L. Allen, M. W. Beijersbergen, R. J. C. Spreeuw, and J. P. Woerdman, “Orbital angular momentum of light and the transformation of Laguerre-Gaussian laser modes,” *Phys. Rev. A*, vol. 45, no. 11, pp. 8185–8189, Jun. 1992, doi: 10.1103/PhysRevA.45.8185.

[5] S. Liu, S. Chen, S. Wen, and H. Luo, “Photonic spin Hall effect: fundamentals and emergent applications,” *Opto-Electronic Sci.*, vol. 1, no. 7, pp. 220007–220007, 2022, doi: 10.29026/oes.2022.220007.

[6] S. Pancharatnam, “Generalized theory of interference and its applications,” *Proc. Indian Acad. Sci. - Sect. A*, vol. 44, no. 6, pp. 398–417, Dec. 1956, doi: 10.1007/BF03046095.

[7] J. D. Jackson, “Electrodynamics, Classical,” in *digital Encyclopedia of Applied Physics*, Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA, 2003. doi: 10.1002/3527600434.eap109.

[8] K. Y. Bliokh and A. Aiello, “Goos-Hänchen and Imbert-Fedorov beam shifts: An overview,” *Journal of Optics (United Kingdom)*, vol. 15, no. 1. IOP Publishing Ltd, 2013. doi: 10.1088/2040-8978/15/1/014001.

[9] K. Y. Bliokh, Y. Gorodetski, V. Kleiner, and E. Hasman, “Coriolis effect in optics: Unified geometric phase and spin-hall effect,” *Phys. Rev. Lett.*, vol. 101, no. 3, Jul. 2008, doi: 10.1103/PhysRevLett.101.030404.

[10] C. Imbert, “Calculation and Experimental Proof of the Transverse Shift Induced by Total Internal Reflection of a Circularly Polarized Light Beam,” *Phys. Rev. D*, vol. 5, no. 4, pp. 787–796, Feb. 1972, doi: 10.1103/PhysRevD.5.787.

[11] T. Zhu *et al.*, “Generalized Spatial Differentiation from the Spin Hall Effect of Light and Its Application in Image Processing of Edge Detection,” *Phys. Rev. Appl.*, vol. 11, no. 3, 2019, doi: 10.1103/PhysRevApplied.11.034043.

[12] M. N. Baitha and K. Kim, “Polarization Switchable Enhanced Photonic Spin Hall Effect for Optical Switch and Filters,” *IEEE Photonics Technol. Lett.*, vol. 34, no. 24, pp. 1329–1332, 2022, doi: 10.1109/LPT.2022.3215763.

[13] X. Ling, W. Xiao, S. Chen, X. Zhou, H. Luo, and L. Zhou, “Revisiting the anomalous spin-Hall effect of light near the Brewster angle,” *Phys. Rev. A*, vol. 103, no. 3, p. 33515, 2021, doi: 10.1103/PhysRevA.103.033515.

[14] R. G. Wan and M. S. Zubairy, “Controlling photonic spin Hall effect based on tunable surface plasmon resonance with an N-type coherent medium,” *Phys. Rev. A*, vol. 101, no. 3, pp. 1–8, 2020, doi: 10.1103/PhysRevA.101.033837.

[15] M. N. Baitha and K. Kim, “Polarization manipulation of giant photonic spin Hall effect using wave-guiding effect,” *J. Appl. Phys.*, vol. 132, no. 5, 2022, doi: 10.1063/5.0100554.

[16] P. V. Kapitanova *et al.*, “Photonic spin Hall effect in hyperbolic metamaterials for polarization-controlled routing of subwavelength modes,” *Nat. Commun.*, vol. 5, 2014, doi: 10.1038/ncomms4226.

[17] Y. He *et al.*, “Controllable photonic spin Hall effect with phase function construction,” *Photonics Res.*, vol. 8, no. 6, p. 963, Jun. 2020, doi: 10.1364/prj.388838.

[18] A. K. Goyal, D. Divyanshu, and Y. Massoud, “Nanophotonic resonator assisted photonic spin Hall enhancement for sensing application,” *Sci. Rep.*, vol. 13, no. 1, pp. 1–9, 2023, doi: 10.1038/s41598-023-36417-5.

[19] F. Wu *et al.*, “Strong polarization-dependent photonic bandgap and giant photonic spin Hall effect in a one-dimensional photonic crystal entirely composed of all-dielectric metamaterials,” *Phys. B Condens. Matter*, vol. 670, no. 293, p. 415348, 2023, doi: 10.1016/j.physb.2023.415348.

[20] W. Zhu and W. She, “Enhanced spin Hall effect of transmitted light through a thin epsilon-near-zero slab,” *Opt. Lett.*, vol. 40, no. 13, p. 2961, Jul. 2015, doi: 10.1364/ol.40.002961.

[21] W. J. M. Kort-Kamp, F. J. Culchac, R. B. Capaz, and F. A. Pinheiro, “Photonic spin Hall effect in bilayer graphene moiré superlattices,” *Phys. Rev. B*, vol. 98, no. 19, pp. 1–7, 2018, doi: 10.1103/PhysRevB.98.195431.

[22] M. Cheng, P. Fu, and S. Chen, “Enhanced and tunable photonic spin Hall effect in metasurface bilayers,” *J. Opt. Soc. Am. B*, vol. 39, no. 1, p. 316, Jan. 2022, doi: 10.1364/josab.443687.

[23] V. Sherawat, R. Bokolia, and R. K. Sinha, “Impact of thermal and refractive index tuning on the bandgap and band-edges of a silicon photonic crystal waveguide with sensing applications,” *Opt. Commun.*, vol. 518, no. April, p. 128348, 2022, doi: 10.1016/j.optcom.2022.128348.

[24] V. Sherawat, R. Bokolia, and R. K. Sinha, “Studies on temperature-dependent bandgap and gap-to-midgap ratio in diamond lattice photonic crystal with biosensing applications,” *Opt. Mater. (Amst).*, vol. 145, p. 114470, Nov. 2023, doi: 10.1016/j.optmat.2023.114470.

[25] J. Cheng, G. Wang, P. Dong, D. Liu, F. Chi, and S. Liu, “Photonic spin Hall effect and terahertz gas sensor via InSb-supported long-range surface plasmon resonance,” *Chinese Phys. B*, vol. 31, no. 1, Jan. 2022, doi: 10.1088/1674-1056/ac192a.

[26] J. Cheng, Y. Xiang, J. Xu, S. Liu, and P. Dong, “Highly Sensitive Refractive Index Sensing Based on Photonic Spin Hall Effect and Its Application on Cancer Detection,” *IEEE Sens. J.*, vol. 22, no. 13, pp. 12754–12760, 2022, doi: 10.1109/JSEN.2022.3177523.

[27] A. K. Sharma, P. Kumar, and Y. K. Prajapati, “Plasmonics-based gas sensor with photonic spin hall effect in broad terahertz frequency range under variable chemical potential of

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:50:46 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3412927

graphene," *Opt. Quantum Electron.*, vol. 54, no. 6, Jun. 2022, doi: 10.1007/s11082-022-03626-7.
[28] X. Zhou, X. Ling, H. Luo, and S. Wen, "Identifying graphene layers via spin Hall effect of light," *Appl. Phys. Lett.*, vol. 101, no. 25, p. 251602, Dec. 2012, doi: 10.1063/1.4772502.
[29] J.-Y. Sui, S. Liao, B. Li, and H.-F. Zhang, "High sensitivity multitasking non-reciprocity sensor using the photonic spin Hall effect," *Opt. Lett.*, vol. 47, no. 23, p. 6065, Dec. 2022, doi: 10.1364/ol.476048.
[30] "AQI Basics." https://www.airnow.gov/aqi/aqi-basics/
[31] D. Fahimi, O. Mahdavipour, J. Sabino, R. M. White, and I. Paprotny, "Vertically-stacked MEMS PM2.5 sensor for wearable applications," *Sensors Actuators, A Phys.*, vol. 299, p. 111569, 2019, doi: 10.1016/j.sna.2019.111569.
[32] P. N. Prasad, *Nanophotonics*. Wiley, 2004. doi: 10.1002/0471670251.
[33] L. M. Zhao and Y. S. Zhou, "Tunable multichannel Photonic spin Hall effect in metal-dielectric-metal waveguide," *Sci. Rep.*, vol. 11, no. 1, pp. 1–6, 2021, doi: 10.1038/s41598-021-93517-w.
[34] O. V. Kotov and Y. E. Lozovik, "Hyperbolic hybrid waves and optical topological transitions in few-layer anisotropic metasurfaces," *Phys. Rev. B*, vol. 100, no. 16, Oct. 2019, doi: 10.1103/PhysRevB.100.165424.
[35] M. Born *et al.*, *Principles of Optics*. Cambridge University Press, 1999. doi: 10.1017/CBO9781139644181.
[36] "AQI Basics | AirNow.gov." https://www.airnow.gov/aqi/aqi-basics/ (accessed Nov. 12, 2023).
[37] L. Lau and J. He, "Investigation into the effect of atmospheric particulate matter (PM2.5 and PM10) concentrations on GPS signals," *Sensors (Switzerland)*, vol. 17, no. 3, 2017, doi: 10.3390/s17030508.
[38] "No Title", [Online]. Available: https://www.airnow.gov/aqi/aqi-basics/
[39] P. Agarwal, K. Kishor, and R. K. Sinha, "Ultrasensitive dual-band terahertz metasurface sensor based on all InSb resonator," *Opt. Commun.*, vol. 522, p. 128667, Nov. 2022, doi: 10.1016/j.optcom.2022.128667.
[40] M. Nejat and N. Nozhat, "Sensing and switching capabilities of a tunable GST-based perfect absorber in near-infrared region," *J. Phys. D. Appl. Phys.*, vol. 53, no. 24, p. 245105, 2020, doi: 10.1088/1361-6463/ab7d6a.
[41] M. R. Forouzeshfard, S. Ghafari, and Z. Vafapour, "Solute concentration sensing in two aqueous solution using an optical metamaterial sensor," *J. Lumin.*, vol. 230, p. 117734, 2021, doi: 10.1016/j.jlumin.2020.117734.
[42] A. Hamouleh-Alipour, M. Forouzeshfard, R. Baghbani, and Z. Vafapour, "Blood Hemoglobin Concentration Sensing by Optical Nano Biosensor-Based Plasmonic Metasurface: A Feasibility Study," *IEEE Trans. Nanotechnol.*, vol. 21, pp. 620–628, 2022, doi: 10.1109/TNANO.2022.3213560.
[43] A. Alipour, A. Farmani, and A. Mir, "High Sensitivity and Tunable Nanoscale Sensor Based on Plasmon-Induced Transparency in Plasmonic Metasurface," *IEEE Sens. J.*, vol. 18, no. 17, pp. 7047–7054, 2018, doi: 10.1109/JSEN.2018.2854882.
[44] A. Keshavarz and Z. Vafapour, "Thermo-optical applications of a novel terahertz semiconductor metamaterial design," *J. Opt. Soc. Am. B*, vol. 36, no. 1, p. 35, 2019, doi: 10.1364/josab.36.000035.

Lokesh Ahlawat is currently serving as PhD Research Scholar in the Department of Applied Physics at Delhi Technological University. He completed his masters from Department of Physics, Maharishi Dayanand University, Rohtak, Haryana, India. He graduated with an Honours degree in Physics from University of Delhi, India. His research areas include Photonic Spin Hall Effect, Topological Photonics Photonic Crystals, Metamaterial and Spin-Orbit Coupling of light.

Kamal Kishor is currently working as Assistant Professor, Applied Physics Department, Delhi Technological university (Formerly known as Delhi college of Engineering, University of Delhi), Delhi since 2011. He received his Ph.D. in "Characterization of Photonic Crystal Fibers and Metamaterials: Theory and Experiments" from the University of Delhi, India. He has more than 17 years of teaching and research experience. He was former member of IEEE. He has also worked as visiting faculty in Department of Physics, Motilal Nehru College, of Delhi, Bhartiya Vidyapeeth College of Engineering, Maharaja Agrasen Institute of Technology, Delhi College of Engineering etc.

He has felicitated with "Research Excellence Award" in 2010 by SPIE-The International Society for Optical Engineering, USA. He is reviewer of many International Journals. His interest area is Optical Communication and Photonics. He is currently working in the area of Optical Fiber Communication System, Photonics Crystal Fibers, Metamaterials based antenna, Sensors, Nano Photonics Devices, etc.

Dr. Kishor is a former member of the International Society of Optical Engineering (SPIE, USA) and a member of the Optical Society of America (OSA, USA).

Monu Nath Baitha is currently serving as a Post-Doctoral Researcher at the School of Mechanical Engineering, Yonsei University in the Republic of Korea. He earned his Ph.D. in Photonic Spin Hall Effect from Yonsei University. Prior to that, he completed his Master's degree in Solid State Materials from the Indian Institute of Technology (IIT) Delhi, India, and his undergraduate degree, B. Tech. in Engineering Physics, from Delhi Technological University, India. Dr. Baitha's research interests encompass spin-orbit interactions of light, topological photonics, metamaterials/metasurfaces, plasmonics, and nanophotonics

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:50:46 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2024.3412927

Ravindra Kumar Sinha completed M.Sc. Physics from IIT Kharagpur in 1984 and Ph.D. in area of Fiber Optics and Optoelectronics in 1989 from IIT Delhi. He did his Post-Doctoral Research at Osaka University for Foreign Studies and Kobe University in Japan and IISc Bangalore, India, during 1989-1991. He has worked at BITS Pilani, REC (now NIT) Hamirpur H.P. and DCE/DTU Delhi. He has established TIFAC-Center of Relevance and Excellence in Fiber Optics and Optical Communication under Technology Vision 2020 Program, Conceptualized and Executed B.Tech. Engineering Physics, M.Tech.(MOCE) and M.Tech. (NST) at DTU Delhi. His major area of research work is Photonic Crystal and Metamaterial-based Nanophotonic Devices for Telecom and Sensing Applications.

He has published over 390 research papers in reputed Journals and Conference Proceedings and 07 book chapters and 03 books, originated 06 patents, supervised 22 sponsored projects, and 21 Doctoral theses. He has served as Director of CSIR-CSIO Chandigarh, CEERI Pilani, and IMTECH Chandigarh. He has mentored over 39 Technology development and transfers to the industry. He is a recipient of the Gold-Skoch Award for Defence Technology 2020, CSIR Technology Award 2018, Fullbright-Nehru Fellowship 2013 as International Educational Administrator, Royal Academy of Engineering (UK) Fellowship 2008, JSPS Fellowship besides several awards for his research work.

Prof. Sinha is the Fellow of the International Society of Optics and Photonics (SPIE), Fellow of IETE, and Fellow of OSA.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 01,2024 at 04:50:46 UTC from IEEE Xplore. Restrictions apply.
© 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Check for updates

Received: 6 April 2024 | Revised: 19 May 2024 | Accepted: 20 May 2024

DOI: 10.1002/pen.26806

RESEARCH ARTICLE

POLYMER ENGINEERING AND SCIENCE WILEY

Physicochemical evaluation of polyvinyl alcohol films crosslinked with saturated and unsaturated dicarboxylic acids: A comparative study

Leela Gautam[1,2] | **Sudhir G. Warkar**[1] | **Manish Jain**[1]

[1]Department of Applied Chemistry, Delhi Technological University, New Delhi, India

[2]Department of Chemistry, Zakir Husain Delhi College (University of Delhi), New Delhi, India

Correspondence
Manish Jain, Department of Applied Chemistry, Delhi Technological University, New Delhi, India.
Email: manishjain@dtu.ac.in

Abstract

In this study, crosslinked films of polyvinyl alcohol (PVA) were prepared using two dicarboxylic acids: saturated succinic acid (SA) and unsaturated maleic acid (MAL) as crosslinkers. The crosslinking capabilities of these acids were compared through physical, chemical, mechanical, morphological, thermal analysis, and swelling measurements to investigate their impact on the physicochemical properties of the resulting films. Attenuated total reflectance fourier transform infrared spectroscopy (FTIR) confirmed the occurrence of an esterification reaction between PVA and the dicarboxylic acids. Based on the degree of crosslinking and thermal properties, it was noted that the crosslinking achieved with maleic acid is more effective for PVA compared with succinic acid. The crosslinking degree of PVA-MAL was determined to be 6.47 times higher than that of the PVA-SA film. In terms of thermal stability, the decomposition temperature for crosslinked PVA-MAL was measured at 365.57°C, whereas PVA-SA exhibited a slightly lower decomposition temperature of 362.58°C. The findings indicate that the PVA-MAL film displays lower crystallinity and a higher contact angle (80.60°) in contrast to PVA-SA (76.0°) and PVA (30.4°). Atomic force microscopy analysis indicated that the PVA-SA film exhibited more surface roughness compared with the smoother PVA-MAL film. Also, crosslinked films displayed enhanced elasticity and resilience in comparison with neat PVA, which may be due to the difference in crosslinking density, disruption in intermolecular hydrogen bonding due to crosslinking, and incorporation of flexible crosslinkers.

Highlights

- PVA was crosslinked using SA and MAL.
- Unsaturated MAL exhibited superior crosslinking density over saturated SA.
- PVA-SA film exhibited a porous surface compared to PVA-MAL film.
- PVA crosslinked with unsaturated MAL showed higher thermal stability than SA.
- Both the films showed enhanced elasticity and resilience compared with neat PVA.

KEYWORDS

carboxylic acids crosslinked PVA, physicochemical properties of polymer, PVA

 © 2024 Society of Plastics Engineers.

1 | INTRODUCTION

Polyvinyl alcohol (PVA) stands out as a highly promising semicrystalline polymer with a range of notable properties, including biocompatibility, water solubility, ready availability, low cost, non-toxicity, and biodegradability.[1–3] It is widely used for the production of thin films in the field of packaging,[4] drug delivery,[5] pervaporation,[6] tissue engineering,[7] and fuel cells.[8] However, PVA films face challenges due to their limited mechanical strength, and their hydrophilic characteristics contribute to poor moisture barrier properties and higher water sensitivity. To improve water resistance and material properties of PVA, a chemical modification that is crosslinking is required.[2,9,10]

Various crosslinkers are commonly utilized for this purpose, including dialdehydes,[11] boric acid,[12] genipin,[13] carboxylic acids,[14] and epoxy compounds.[15] Among them, carboxylic acids have emerged as promising crosslinkers due to their easy availability, environmental friendliness, cost-effectiveness, non-toxicity, biodegradability, and non-accumulative nature.[16,17] Furthermore, the crosslinking of PVA with carboxylic acids involves an esterification process, with water being the primary by-product. This aspect contributes to the overall environmental friendliness of both the crosslinkers and the process itself, as the reactions generate no hazardous or toxic by-products harmful to the environment. A recent review provides a comprehensive overview of the applications and properties associated with PVA crosslinked using carboxylic acids.[18]

While significant researches have been dedicated to the crosslinking of PVA using carboxylic acids, only a limited number of studies have explored the influence of their molecular structure on physicochemical properties.[19–22] These investigations revealed the crucial role played by the molecular structure of carboxylic acids in shaping the degree of crosslinking with PVA and, consequently, the resulting film properties. A comparative study involving suberic acid as aliphatic molecule and terephthalic acid as aromatic molecule revealed the greater compatibility of aliphatic molecules for crosslinking compared with aromatic ones and which in turn showed the elevated tensile strength and thermal properties.[9] In another study, the impact of carboxylic group positioning in the aliphatic chain was studied and it was found that the trans position of carboxyl group in fumaric acid exhibited enhanced crosslinking compared with maleic acid due to its ability to easily get fit in between the polymeric chains, which further influenced the pervaporation process.[22,23] Similarly, an examination of phthalic acid, isophthalic acid, and terephthalic acid elucidated that the para position in terephthalic acid proved most effective for crosslinking, whereas the ortho position was summarized as less suitable.[23] Furthermore, the research highlighted the role of carbon chain length in crosslinking, with succinic acid and adipic acid in the dicarboxylic series demonstrating that longer carbon chains create more space within the PVA matrix, facilitating increased water permeability.[21] These findings collectively emphasize the importance of considering molecular structure, carboxylic group positioning, and carbon chain length when selecting crosslinkers for specific applications to achieve desired properties in PVA films. While independent reports exist in the published literature regarding the properties of PVA crosslinked with saturated and unsaturated crosslinkers, but there is an absence of a comprehensive comparative study conducted under uniform experimental conditions.[24–27] Such a study is crucial for understanding the influence of saturated and unsaturated crosslinking agents on the physicochemical properties of PVA.

This experimental work is therefore intended to fill this gap and specifically aims to elucidate the impact of single and double bonds present in the crosslinking agents. For this investigation, two dicarboxylic acids, succinic acid (SA), and maleic acid (MAL) were chosen. Succinic acid features a saturated, linear chain containing four carbon, between, whereas maleic acid is characterized by an unsaturated chain containing four carbon atoms containing a double bond between two carboxylic acid groups. The study involves a comparative assessment of the chemical structure, hydrophilicity, swelling degree, thermal, tensile, and morphological properties of crosslinked PVA films using various techniques such as attenuated total reflectance (ATR)-FTIR, x-ray diffraction (XRD), scanning electron microscopy (SEM), atomic force microscopy (AFM), differential scanning calorimetry (DSC), tensile, swelling measurements, and sessile drop contact angle analysis. The detailed findings of these investigations are presented and correlated with the structure of the dicarboxylic acids.

2 | MATERIALS AND METHODS

2.1 | Materials

PVA of M.W. = 145,000, fully hydrolyzed was ordered from Sigma-Aldrich, USA. Dicarboxylic acids namely SA and MAL were purchased from Loba Chemie and used as received without any purification. Deionized water was used for all experiments. The molecular structure of acids is depicted in Figure 2.

2.2 | Methodology

The crosslinked PVA films were synthesized using the solution casting method, illustrated in Figure 1.

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

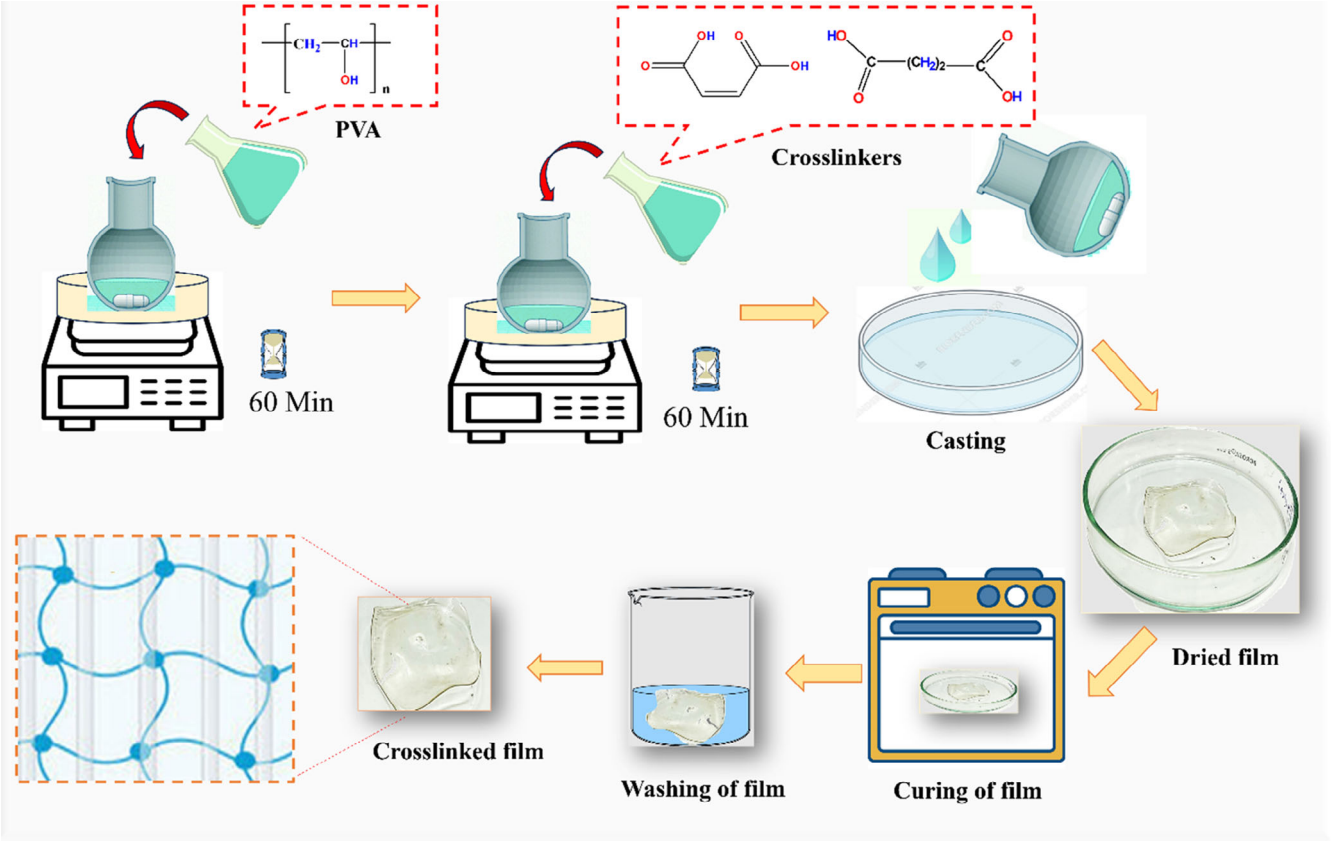


FIGURE 1 Schematic diagram of preparation of crosslinked polyvinyl alcohol (PVA) film.

TABLE 1 Compositions and formulation of crosslinked polyvinyl alcohol (PVA) films.

S. No.	Designation	PVA (mL)	Succinic acid (g)	Maleic acid (g)	Curing temperature (°C)	Curing time (Min)
1.	PVA-SA	20	0.48	-	100	90
2.	PVA-MAL	20	-	0.48	100	90

Abbreviations: MAL, maleic acid; SA, succinic acid.

Initially, an 8% w/v PVA solution was prepared by dissolving PVA in water under stirring at 80°C for approximately 1 h, ensuring the formation of a homogeneous and transparent solution. In the next step, 20 mL of the PVA solution was taken in a round bottom flask attached to a condenser, and 30% by weight of succinic acid and maleic acid, respectively, were added. The resulting solution was stirred in a water bath for 1 h at 80°C to achieve complete homogenization. The clear solution obtained was then poured into molds made of flat glass sheets for the complete evaporation of water under environmental conditions, forming the film. This process was repeated for each acid. The dried films were peeled off and subjected to curing at 100°C for 90 min to promote the crosslinking reaction. To eliminate any unreacted crosslinker, the cured films were washed several times by placing them in distilled water to remove extracted acids. The washed crosslinked films were air-dried at room temperature and subsequently stored in sealed plastic bags for further research. The films were named PVA-MAL and PVA-SA, and their composition and formulation details are provided in Table 1.

3 | CHARACTERIZATION OF THE CROSSLINKED FILMS

3.1 | ATR—FTIR spectroscopy

To confirm the presence of characteristic functional groups, ATR-FTIR analysis was conducted using the Brukeroptik Spectrum instrument. The FTIR spectra of the samples were recorded at resolution 4 cm^{-1}, and the number of scans was 48 in the range of 4000–500 cm^{-1}.

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The obtained values were processed using Origin Lab software to generate and visualize the spectra.

3.2 | Swelling percentage and crosslinking density

Crosslinked PVA films, varying in dicarboxylic acids, were cut into 1.5 cm × 1.5 cm pieces, weighed, and submerged in deionized water (pH ∼ 7.0–7.2) at room temperature. At different intervals, the swollen films were removed from the water, weighed after gently wiping off excess water with filter paper, and promptly placed back into the water. Initially, measurements were taken hourly for the first 5 h, followed by a final measurement at the 24-h mark. The films were left in water for several additional days and were found to be stable, with the swelling percentage remaining constant as observed at the 24-h measurement. Therefore, we reported the values up to 24 h, considering it as the equilibrium value. The degree of swelling was determined using the following Equation (1).

$$\%S = \frac{W_S - W_D}{W_D} \times 100, \tag{1}$$

where, S, W_S, and W_D are the degree of swelling in percent, the weight of the films in the swollen state, and the weight of the film in dry states, respectively. The above study was performed in triplicate.

However, the swelling characteristics of the crosslinked films can be understood more by calculating parameters such as the number average molecular mass between crosslinks (M_c), crosslink density (ρ), and mesh size (ξ). To predict M_c, the Flory–Rehner equation provided below was used.[28–30]

$$\frac{1}{M_c} = -\frac{\frac{v}{V}[\ln 1 - \varphi] + \varphi + \chi\varphi^2}{\varphi^{1/3} - \frac{1}{2}\varphi}. \tag{2}$$

Here, V represents the molar volume of water ≃molar mass of water (18.1 g/mol). The Flory Huggins polymer–water interaction parameter, denoted as χ, ≃0.498. The specific volume of the polymer, v, is 0.788 cm^3/g. The polymer volume fraction in the swollen state, denoted as φ, was estimated from equilibrium swelling results using the provided relation.[29]

$$\varphi = \left[1 + \frac{d_p}{d_s}\left(\frac{m_a}{m_b}\right) - \frac{d_p}{d_s}\right]^{-1}. \tag{3}$$

To determine φ using Equation (3), the calculation involved the following parameters: mass of the film before swelling (m_b), the mass of the film after swelling (m_a), the density of the hydrogel (d_p), and the density of the solvent (d_s). The density of the films was calculated using Equation (4).

$$d_p = \frac{W}{S \times X} \tag{4}$$

In the given context, X denotes the average thickness of the film, S represents the cross-sectional area, and W corresponds to the weight of the film. The density (ρ) of the films was determined using the following equation.

$$\rho = \frac{1}{v \times M_c}. \tag{5}$$

The mesh size (ξ) was calculated using the Flory characteristic ratio of the polymer C by the following equation[28]:

$$\xi = \varphi^{\frac{-1}{3}}\left[c_n\left(\frac{M_c}{M_r}\right)\right]^{\frac{-1}{2}}.l \tag{6}$$

Here, C_n is the Flory ratio (with a value of 8.3), l represents the bond angle between the C—C bond (measured at 1.54 Å), M_r is the molecular weight of repeating units of PVA, and M_c represents the molecular weight of crosslinks.

3.3 | Water contact angle measurement

The hydrophilicity of the PVA crosslinked films was investigated by a Kruss advance 1.6.2.0 sessile drop contact angle Goniometer at ambient temperature. A deionized water droplet of 2 μL was dropped by a microsyringe onto the top surface of the films. For each film, a minimum of 5 points were measured to get the average water contact angle measurement (WCA).

3.4 | XRD analysis

To study the crystalline structure of the PVA crosslinked films, XRD patterns of all crosslinked films were recorded using the Rigaku Ultima IV diffractometer (Japan) [λ(Cu–Kα) = 1.54 A]. Scattered radiations were recorded in the range of 2θ ranging from 5° to 50° at a rate of 4 degrees/min.

3.5 | Field emission scanning electron microscopy

To observe the change in the morphology of the surface of the PVA, crosslinked films were examined using FEI Quanta 200 Field emission scanning electron microscopy

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

(FESEM), Netherlands. The films were coated with gold–palladium and observed at a voltage of 5.0 kV.

3.6 | Atomic force microscopy

AFM images were recorded using asylum research MFP3D-BIO atomic force microscope to observe the 3-D image and to estimate the roughness of the films.

3.7 | Thermal gravimetric analysis

To observe the thermal stability and decomposition temperature of neat PVA and crosslinked PVA samples, thermal gravimetric analysis (TGA) and DTGA were recorded using PerkinElmer TGA 4000 with an N_2 atmosphere in the range of 30–700°C at the rate of 10°C/min.

3.8 | Differential scanning calorimetry

DSC analysis was performed to examine the change in heat evolved and melting point in the PVA crosslinked films. The heat flow measurement was recorded using the PerkinElmer model DSC8000 analyzer. The samples of nearly 10 mg mass of pure PVA and crosslinked PVA films were heated from 30 to 300°C at 10°C/min.

3.9 | Tensile strength

The tensile tests of the PVA and crosslinked films were performed on Zwick Roell Z010 according to ASTM D638 and ASTMD790, respectively. Three samples were tested for each set, and the mean value was reported.

3.10 | Rheological studies

Dynamic rheological studies of neat PVA and crosslinked PVA films were carried out using Anton Paar MCR-702, rheometer in torsional setup mode. Frequency-dependent data were collected at room temperature and a fixed strain ($\gamma = 0.3\%$) over a frequency range from 0.01 to 100 Hz.

4 | RESULTS AND DISCUSSION

4.1 | Formation of crosslinked films

Crosslinked PVA films, prepared at a 30% w/w concentration of SA and MAL, a curing temperature of 100°C, and a curing time of 90 min. The concentration, curing time, and temperature were determined based on preliminary studies, revealing that temperatures exceeding 100°C and curing times beyond 90 min resulted in the appearance of yellow color on the films. PVA films crosslinked with SA and MAL were obtained through an esterification crosslinking mechanism. Being difunctional acids, both SA and MAL may react with the hydroxyl groups of PVA chains, and this results in intermolecular-type and intramolecular-type ester linkages among them. Maleic acid or cis-butenedionic acid (MAL) consists of four carbon atoms. However, its molecular structure includes a carbon–carbon double bond (C═C), introducing rigidity in terms of group rotations around the C—C axis. This double bond restricts the rotation of the COOH groups about the C—C bond, restricting certain molecular motions.[31] In comparison, SA comprises only single bonds and has free rotations around the C—C axis, resulting in a comparatively more flexible structure than MAL. The best possible interaction of SA and MAL based on the information available in the literature is depicted in Figure 2. In our current study, homogeneous solutions of PVA and acids were dried at room temperature. During drying, the reactive functional groups of PVA and acid molecules aligned, potentially forming H-bonds and some of the ester bonds, as previously proposed by Larrañeta et al. and Liu et al.[29,32] To enhance the crosslinking reaction, the films were further cured by heating them in an oven at 100°C for 90 min. The supplied thermal energy during crosslinking facilitated the formation of more ester bonds between the previously aligned COOH and OH groups. Additionally, water evaporation occurred, promoting the esterification reaction in the forward direction and lowering the pH of the prepared films, thereby catalyzing esterification.[32] Washing the cured films in water helped leach out most of the unreacted acid, leaving behind ester-bonded and strongly H-bonded acids. The prepared films were further characterized by ATR-FTIR to prove the formation of ester bonds, discussed later.

4.2 | ATR—FTIR spectroscopy

The ATR-FTIR spectra and data of neat PVA, SA, MAL, and crosslinked PVA films are shown in Figure 3. From the graph, it is evident that when SA and MAL were used as crosslinkers with PVA, new peaks emerged in the FTIR spectrum at 1730 cm^{-1} for PVA-SA and 1720 cm^{-1} for PVA-MAL films. These peaks indicate that an esterification reaction has occurred between PVA and both acids. The peaks correspond to the C═O stretching frequency, confirming ester formation as a result of crosslinking.

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

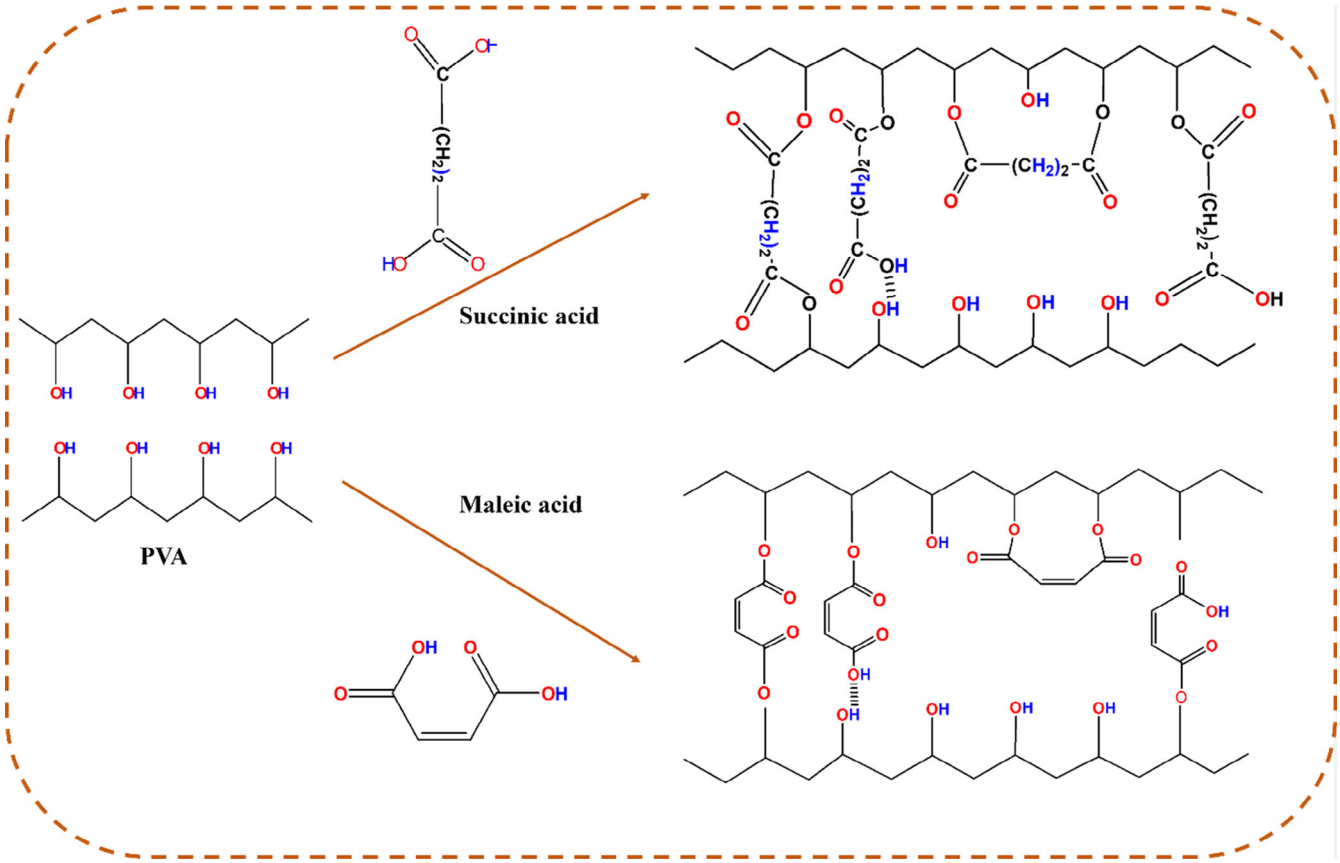


FIGURE 2 The molecular structures of the polyvinyl alcohol (PVA), crosslinkers, and schematics illustrating the modes of crosslinking with PVA.

In PVA-MAL, the C═O stretching is observed at slightly lower frequency, influenced by the conjugation with C═C bond.[33]

As the crosslinking process does not alter the existing methylene groups (1420 cm^{-1}) in the crosslinked films, the intensity of the methylene group peak remains constant. Therefore, the extent of crosslinking can be quantitatively determined using the band index method, which involves calculating the area ratio of the C═O (ester bond) peak to the CH_2 (methylene group) peak.[33] The results revealed that the ratio for the PVA-MAL film (0.63) film is higher than the PVA-SA (0.38). This observation suggests an increment in the carboxyl group content and further indicates more crosslinking in the PVA-MAL film. This could be due to the higher solubility of maleic acid than succinic acid and may be due to the cis orientation of carboxyl groups.[34]

4.3 | Swelling analysis

The swelling behavior of PVA is primarily influenced by the presence of available free OH groups that attract water molecules and the ease of diffusion of water molecules in the space present between the chains. Crosslinking of PVA with dicarboxylic acids, known for enhancing water resistance, involves the consumption of OH groups in esterification reactions. Additionally, crosslinking covalently binds PVA chains, reducing the available space for water molecules to come inside. Saturated and unsaturated dicarboxylic acids, when employed as crosslinkers, exhibit varying effects on the swelling behavior of PVA, as depicted in Figure 4. The intermolecular or intramolecular crosslinking generates numerous succinate and maleate ester groups, leading to a reduction in swelling percentage. Notably, the reduction in swelling percentage was found more pronounced in PVA-MAL compared with PVA-SA film. This may be due to the presence of double bonds present in MAL resulting in a more rigid structure of the crosslinked film and anticipated to allow less amount of water in comparison with the relatively flexible structure of SA molecule and may provide more space in between the PVA chains.

To go deeper into the analysis, the equilibrium swelling data were utilized to calculate the essential network parameters, including the average molecular weight between consecutive crosslinks (M_c), crosslinking density (ρ), and mesh size (ξ), as presented in Table 2.

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

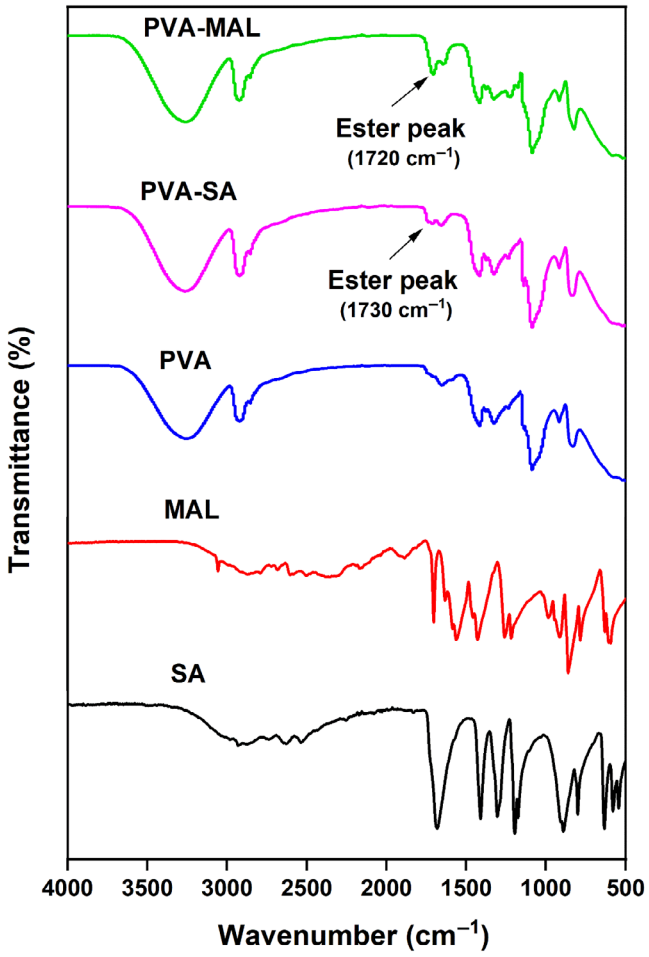


FIGURE 3 Fourier transform infrared spectroscopy (FTIR) analysis of the succinic acid (SA); maleic acid (MAL); polyvinyl alcohol (PVA); SA crosslinked film (PVA-SA); and MAL crosslinked film (PVA-MAL).

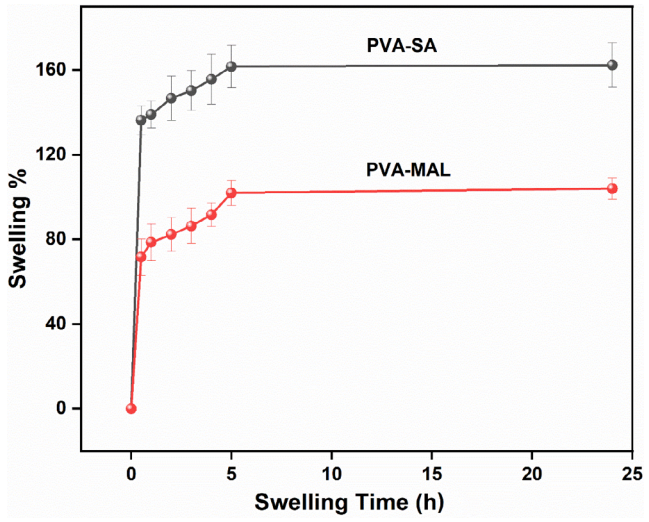


FIGURE 4 Swelling analysis of the dicarboxylic acids crosslinked polyvinyl alcohol (PVA) films; PVA-succinic acid (SA) and PVA-maleic acid (MAL).

The crosslinking density is particularly crucial in understanding swelling behavior, as a lower crosslinking density indicates greater distances between crosslinking points. Additionally, the mesh size provides insights into the available space for water molecules to permeate. As can be seen in Table 2, the average molecular weight between crosslinks and mesh size values was found proportional to the swelling capacity of the film. The M_c, ρ, and ξ values obtained for PVA-SA films were found aligned with swelling results.

4.4 | Water contact angle measurement

The WCA is an effective method for assessing the hydrophilicity/hydrophobicity of a surface. Figure 5 illustrates a reduction in the WCA of crosslinked films compared with neat PVA. The corresponding images reveal water droplets spreading differently on the film surfaces. PVA enhances water adhesion due to its polar OH groups, resulting in a lower contact angle. Crosslinking with SA and MAL acids removes numerous polar OH groups from the PVA chain, leading to higher contact angles.[22] The maximum contact angles for PVA, PVA-SA, and PVA-MAL were observed to be 30.40, 76.0, and 80.60, respectively. Through crosslinking, PVA films transition from a highly hydrophilic state to a moderately hydrophobic state.[35] These results demonstrated that PVA-MAL has the maximum contact angle, showing more effective crosslinking than PVA-SA. The above results are in agreement with FTIR and swelling results.

4.5 | XRD analysis

The XRD graphs of crosslinked films, namely PVA-SA and PVA-MAL were examined and compared with the XRD graph of pure PVA film Figure 6. PVA is recognized as a semicrystalline polymer, featuring both a crystalline phase and an amorphous matrix. The predominant peak observed at 19.28 in the XRD graph of PVA is attributed to its ordered arrangement and crystalline phase. Upon crosslinking, the characteristic peak intensity of PVA at 19.28° was noted to decrease in both PVA-SA and PVA-MAL films. This reduction in peak intensity signifies a decrease in the crystallinity of PVA after crosslinking.[23] The decrease in crystallinity is attributed to the incorporation of large dicarboxylic acid moieties into the structure through interactions with free hydroxyl groups in PVA, resulting in the formation of ester bonds. These changes not only indicate modifications in the crystalline phase of PVA but also suggest increasing molecular disarrangement in the PVA matrix. The relative percent

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

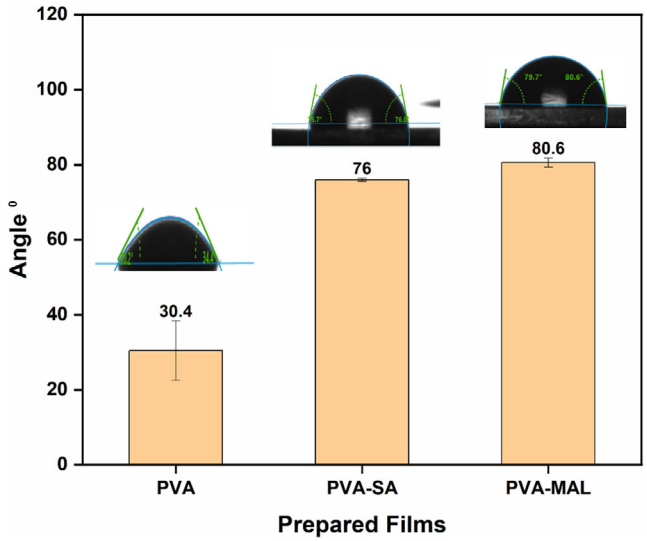


FIGURE 5 Variation of contact angle of polyvinyl alcohol (PVA) and PVA crosslinked films with succinic acid (PVA-SA) and malonic acid (PVA-MAL).

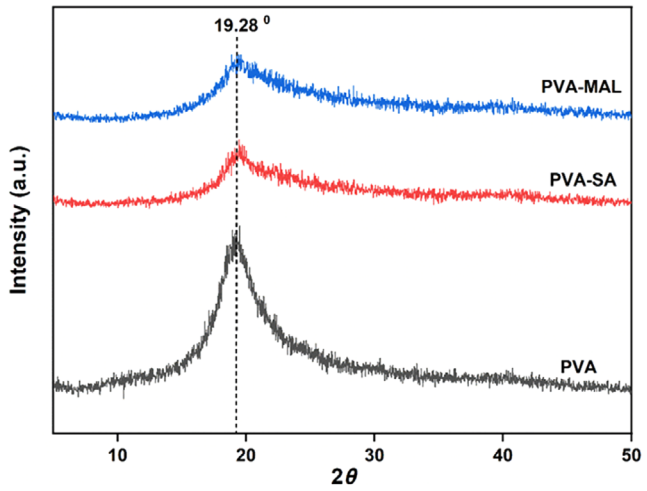


FIGURE 6 X-ray diffraction diffractograms of polyvinyl alcohol (PVA) and PVA crosslinked films with succinic acid (PVA-SA) and maleic acid (PVA-MAL).

TABLE 2 Network parameters of the crosslinked films.

Crosslinked film	φ	M_c (g/mL)	ρ (mol/cm^3)	ξ (Å)
PVA-SA	0.30	980.39	1.29×10^{-3}	95.29
PVA-MAL	0.36	151.97	8.35×10^{-3}	16.39

Abbreviations: MAL, maleic acid; PVA, polyvinyl alcohol; SA, succinic acid.

crystallinity of the crosslinked films was determined with respect to the 19.28 peak of PVA using the following equation.[21]

$$x\% = \frac{Ac}{Ac + A_A} \times 100,$$

where A_C is the crystalline area and A_A is the amorphous area. It was observed that the crystallinity of PVA-MAL (47%) film is lower compared with both PVA-SA (49%) and pure PVA (56%) films. These results regarding crystallinity are in agreement with the outcomes obtained from ATR-FTIR analysis, providing further support to the findings of the study.

4.6 | Field emission scanning electron microscopy

The surface morphology films crosslinked with succinic acid and maleic acid were investigated using FESEM images. Figure 7 illustrates the morphological distinctions between PVA-SA and PVA-MAL films. Notably, the PVA-MAL film displayed a dense structure and a higher level of uniformity and appeared homogenous and smooth which is somewhat similar to the dense layer morphology, suggesting effective dispersion of the crosslinker within the PVA matrix. By contrast, the surface of the PVA-SA film appeared rough, shrunk, and exhibited some pores. This observation signifies that, when subjected to identical processing conditions, the film morphology is significantly influenced by the structure of the crosslinker. The compact structure observed in the PVA-MAL film may be attributed to the effective interlocking of polymer chains, facilitated by the presence of a double bond in MAL.

4.7 | Atomic force microscopy

The 3-D AFM images of PVA and its crosslinked films, that is, PVA-SA and PVA-MAL were analyzed to characterize their surface properties, including roughness and height (Figure 8). The obtained data clearly reveal that the roughness parameters of crosslinked PVA films are greater than the neat PVA film. In the case of PVA, the AFM image reveals a smooth surface with low height (200 nm) and moderate average roughness. Upon crosslinking with SA and MAL, notable changes occurred. The height increases to 760 nm and 590 nm for PVA-SA and PVA-MAL films, respectively. Additionally, there is an increase in roughness values, confirming the occurrence of a chemical reaction during the crosslinking process.[36,37] For PVA, PVA-SA, and PVA-MAL, the increased R_a (average surface roughness) value follows the order of PVA (53 nm) < PVA-MAL (196 nm) < PVA-SA (245 nm). Notably, PVA-SA exhibits significantly higher roughness parameters compared with PVA-MAL also the surface displays distinct humps and grooves, suggesting the formation of voids. The above result matches the findings of SEM analysis.

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

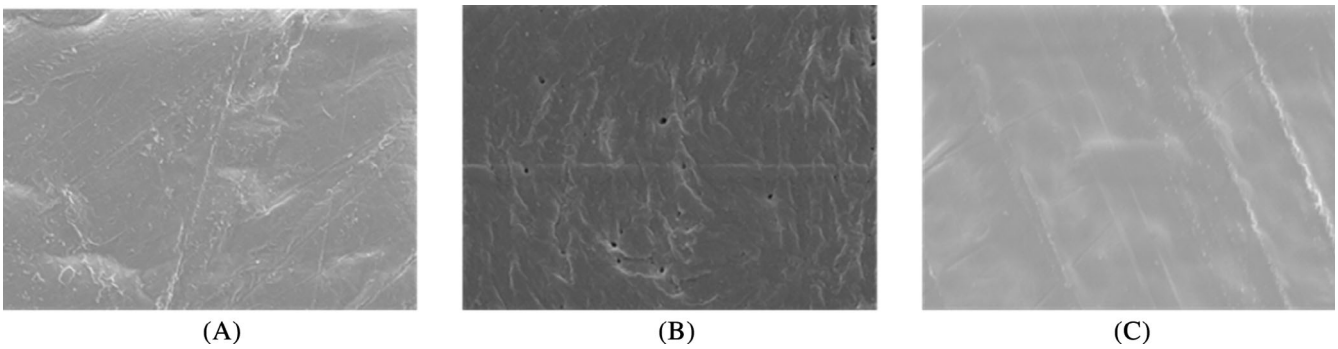
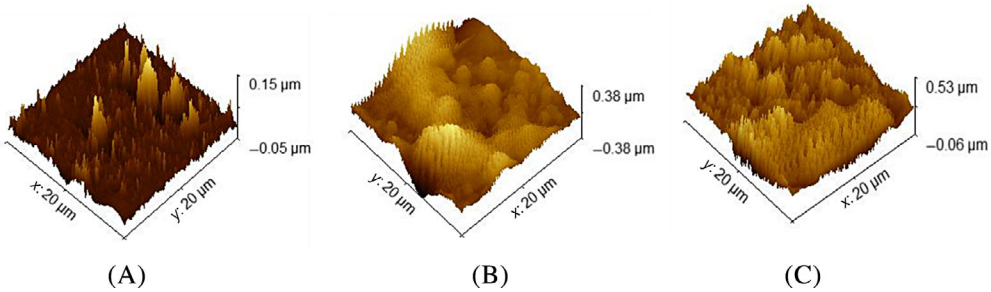


FIGURE 7 Surface field emission scanning electron microscopy images polyvinyl alcohol (PVA) crosslinked films: (A) PVA; (B) succinic acid (PVA-SA); (C) maleic acid (PVA-MAL).

FIGURE 8 Atomic force microscopy images of polyvinyl alcohol (PVA) and PVA crosslinked films: (A) PVA; (B) succinic acid crosslinked film (PVA–SA); (C) maleic acid crosslinked film (PVA-MAL).



Also, the obtained roughness values can be further used to validate the swelling analysis. It was noted that the crosslinked films treated with MAL exhibited relatively less roughness and consequently showed less swelling. This consistency in results could be related to the fewer surface irregularities that restrict water absorption, facilitating a more controlled swelling process.[36,38]

4.8 | Thermal gravimetric analysis

The TGA and DTG thermogram of the pure PVA and crosslinked films are shown in Figure 9A.

All TGA graphs show three-stage degradation. For PVA, initial mass loss is observed due to the moisture evaporation (30–200°C, 8% loss). The main degradation starts from the second stage (263–498°C, 85.5% loss) due to the removal of hydroxyl groups in the form of water and the formation of polyene macromolecules.[5] The third stage (502–689°C, 5%) is because of the overall decomposition of the backbone of PVA.[35] For crosslinked films, the first stage that corresponds to a minimal mass loss is due to the removal of moisture and the second stage at elevated temperatures (269–481°C), which may be due to the breakage of ester linkage or decarboxylation of free carboxylic acids present in the film. The third stage should be the result of the degradation of the backbone of crosslinked films.[35] Additionally, there is a difference in decomposition rates of PVA-SA and PVA-MAL films, which can be attributed to the different crosslinking degrees. At 350°C, PVA-MAL film retains 72% of its original weight whereas PVA-SA and PVA have 71% and 52%. The PVA-MAL film also showed a slightly higher residue left at 700°C than the other two films. In the DTG analysis, Figure 9B, it was observed that the PVA-MAL film exhibited a higher maximum decomposition temperature compared with PVA-SA and PVA. The decomposition temperature for PVA film was recorded at 344.78°C, whereas for PVA-SA and PVA-MAL, the decomposition temperatures were noted at 362.58 and 365.57°C, respectively. These findings indicate that the PVA-MAL films demonstrate enhanced thermal stability compared with the PVA-SA film.

4.9 | Differential scanning calorimetry

The DSC curves of the pure PVA and crosslinked films are shown in Figure 9C.

In the DSC curves, the endothermic peak around 145°C is associated with the evaporation of bound water (T_e), representing the energy needed to vaporize water bound within the films. Another endothermic peak at approximately 229°C corresponds to the crystalline melting (T_m) of the PVA film, with the peak area (ΔH_m) in the melting region indicating the crystallinity of the

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

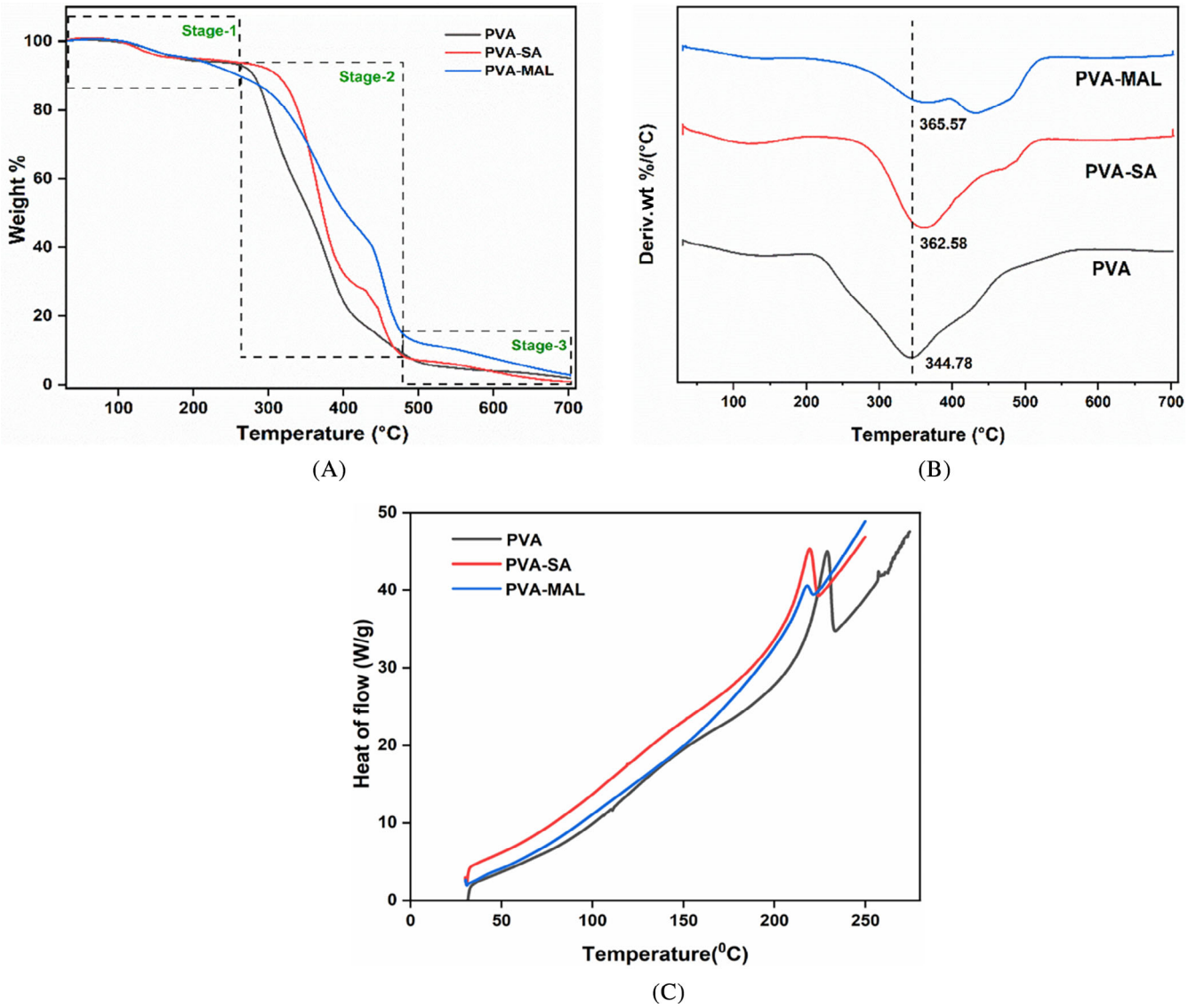


FIGURE 9 (A) Thermal gravimetric analysis, (B) differential thermogravimetry analysis, (C) differential scanning calorimetry graphs of polyvinyl alcohol (PVA) and PVA crosslinked films. MAL, maleic acid; SA, succinic acid.

films. The addition of SA and MAL results in a reduction of the T_e value, indicative of a chemical crosslinking effect.[39,40] This implies that PVA films crosslinked with SA and MAL exhibit reduced hydrophilicity compared with pure PVA. Crosslinking induces disruption in the crystalline region, leading to an increase in the amorphous region and, consequently, a decrease in T_m.[41] Pure PVA, being semicrystalline with a more ordered structure, exhibits a higher melting temperature and a sharp endothermic curve. By contrast, PVA-MAL and PVA-SA show lower melting temperatures and broadened peaks, confirming effective crosslinking. The melting temperature (T_m) and ΔH_m values of the crosslinked PVA films are lower than those of pure PVA, following the order: PVA (229°C, 141.68 J/g) > PVA-SA (211°C, 129.77 J/g) > PVA-MAL (209°C, 21.811 J/g). This indicates that the addition of a crosslinker has diminished the overall crystallinity of the films,[40] matching with the previously discussed XRD results. Consequently, the discussion leads to the conclusion that the PVA-MAL film exhibits the highest thermal stability among the three, consistent with the trends observed in swelling and FTIR results.

4.10 | Tensile properties

The typical tensile stress–strain curves of the PVA, PVA-SA, and PVA-MAL films are shown in Figure 10. Neat PVA film exhibited higher tensile strength than PVA-SA and PVA-MAL films. The tensile strength values follow the order PVA-MAL (30.41 ± 9.02 MPa) < PVA-SA (35.85 ± 3.33 MPa) < PVA (42.77 ± 3.145 MPa).

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

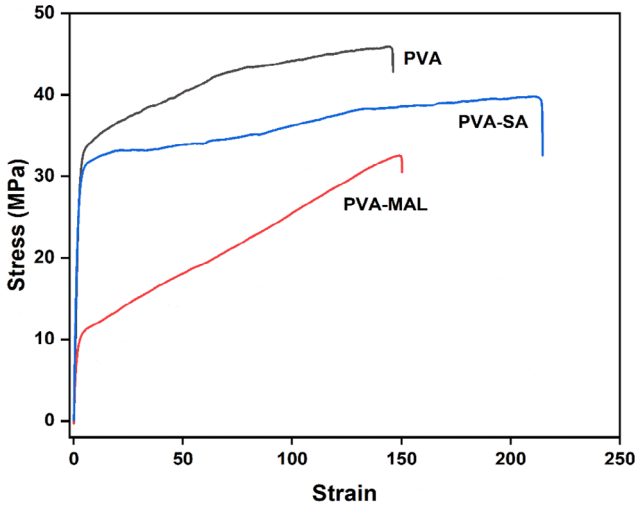


FIGURE 10 Typical stress–strain curve of polyvinyl alcohol (PVA) and PVA crosslinked films: succinic acid crosslinked film (PVA-SA); and maleic acid crosslinked film (PVA-MAL).

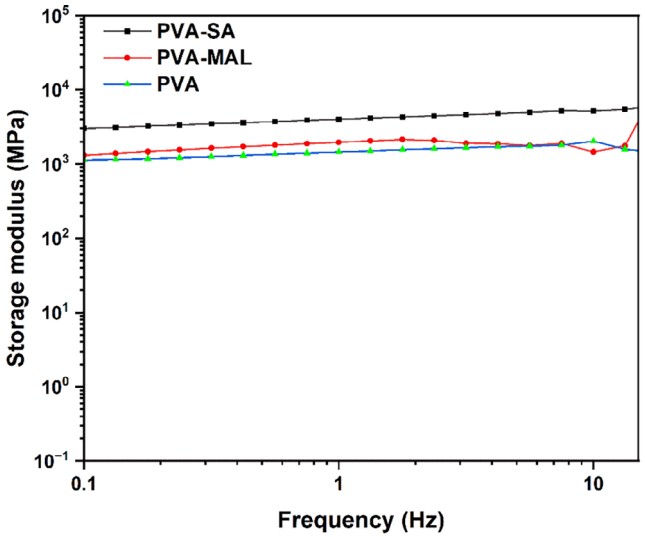


FIGURE 11 Storage modulus of polyvinyl alcohol (PVA) and PVA crosslinked films: Succinic acid crosslinked film (PVA-SA); and maleic acid crosslinked film (PVA-MAL).

The higher tensile strength observed in PVA can be attributed to its crystalline phase and inherent rigidity resulting from strong hydrogen bonding. The decline in tensile strength in crosslinked films may be linked to a reduction in crystallinity and rigidity caused by the introduction of SA and MAL as crosslinkers.[32] The inclusion of crosslinkers, significantly weakens hydrogen bonding and interrupts crystallinity, resulting in lower rigidity and the formation of amorphous regions, which further leads to the reduction in the tensile strength.[42] The higher recorded crosslinking density in PVA-MAL films may be used to explain their lower tensile strength compared with PVA-SA films. Similar results have been reported in some of the studies.[19] Notably, despite the reduction in tensile strength, the crosslinked films exhibit greater elongation when compared to the neat PVA, with % values of 145.06% ± 2.93%, 210.07% ± 4.53%, and 151.41% ± 1.09% for PVA, PVA-SA, and PVA-MAL, respectively. The increased elongation may be associated with the flexible structure of four carbons present in both crosslinkers. However, MAL, with its double bond, promotes more interlocking of chains, resulting in lower elongation for PVA-MAL compared with PVA-SA. It can be concluded that both the crosslinking density and flexibility of the crosslinker affect the mechanical properties of the crosslinked films.

4.11 | Rheological studies

Rheological measurements were carried out to determine the viscoelastic behavior of the PVA, PVA-SA, and PVA-MAL films. Frequency-dependent viscoelastic response of storage modulus (G') is shown in Figure 11. It is well known that the storage modulus of crosslinked polymeric material depends on factors like the extent of crosslinking and crystallinity.[32,42] When a crosslinker is introduced into a semicrystalline polymer, such as the PVA chain, it induces crosslinking but also affects the existing crystallinity. The obtained findings can be explained by focusing on the above two competitive factors during the crosslinking process. From the graph, it is evident that the inclusion of a crosslinker has a significant impact on the storage modulus values of PVA-SA and PVA-MAL films, which are both higher than those of neat PVA. This confirms that effective crosslinking has occurred in both types of films.

Interestingly, the storage modulus (G') of the PVA-SA film is higher than that of the PVA-MAL film, despite the PVA-SA film having a lower crosslinking density. This discrepancy can be attributed to the higher crystallinity observed in the PVA-SA film compared with the PVA-MAL film. The DSC and XRD findings support this observation by showing a greater reduction in crystallinity in the PVA-MAL film relative to the PVA-SA film. The more significant reduction in crystallinity in the PVA-MAL film effectively offsets the impact of its higher crosslinking density, resulting in a lower storage modulus when compared to the PVA-SA film.

5 | CONCLUSION

The study aimed to compare the effect of saturation (single bond) of SA and unsaturation (double bond) of maleic acid as crosslinkers on the properties of PVA

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

films. The crosslinked films were synthesized without the use of any catalyst, highlighting a catalyst-free approach. The calculated crosslinking density of the PVA-MAL film was noticed to be 6.47 times higher than that of the PVA-SA film. Consequently, PVA-MAL films exhibited minimal swelling. Morphological analysis further revealed that PVA-MAL films displayed smoother surfaces than PVA-SA films, which gives an idea of better dispersion of MAL over SA as crosslinker. In thermal analysis, the PVA-MAL film demonstrated superior properties compared with the PVA-SA film. The maximum decomposition temperature recorded for maleic acid crosslinked PVA samples was 365.57°C, higher than succinic acid crosslinked PVA samples (362.58°C) and neat PVA (344.78°C).

XRD results indicated that crosslinking led to a decrease in the overall crystallinity of PVA films, resulting in lower percentages of crystallinity for both crosslinked films, whereas from DSC results, it was seen that PVA-MAL film has less crystallinity % than PVA-SA film. Additionally, a more significant reduction in hydrophilicity was observed in the PVA-MAL film compared with both PVA-SA and neat PVA films. Also, crosslinked films displayed enhanced elasticity and resilience in comparison with neat PVA, with the order of elongation being PVA-SA (210.07% ± 4.53%) > PVA-MAL (151.41% ± 1.09%) > PVA (145.06% ± 2.93%), which may be due to the difference in crosslinking density, disruption in intermolecular hydrogen bonding due to crosslinking, and incorporation of flexible crosslinkers.

Succinic acid and maleic acid crosslinked PVA films can be used for different applications such as food packaging, reverse osmosis membranes for waste water treatment, hydrophilic pervaporation membranes for hydrophobic/hydrophilic separation, fuel cell, wound dressing, drug delivery, etc.[18] The collective findings of this study emphasize that crosslinkers with double bonds can create dense networks, resulting in less swelling. Additionally, double bonds improve the thermal behavior and surface properties of PVA more effectively than single bond containing crosslinkers. These considerations are crucial for tailoring PVA films to achieve specific properties for diverse applications of PVA films but also contribute significantly to the comprehension of green crosslinking processes.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Sudhir G. Warkar https://orcid.org/0000-0001-7305-282X
Manish Jain https://orcid.org/0000-0001-8174-4754

REFERENCES

1. Jose J, Al-Harthi MA. Citric acid crosslinking of poly(vinyl alcohol)/starch/graphene nanocomposites for superior properties. *Iran Polym J*. 2017;26:579-587.
2. Ávila-Salas F, Marican A, Pinochet S, et al. Film dressings based on hydrogels: simultaneous and sustained-release of bioactive compounds with wound healing properties. *Pharmaceutics*. 2019;11:447. doi:10.3390/pharmaceutics11090447
3. Rizwana N, Maslekar N, Chatterjee K, Yao Y, Agarwal V, Nune M. Dual crosslinked antioxidant mixture of poly (vinyl alcohol) and cerium oxide nanoparticles as a bioink for 3D bioprinting. 2023. doi:10.1021/acsanm.3c02962
4. Alaş MÖ, Doğan G, Yalcin MS, Ozdemir S, Genç R. Multicolor emitting carbon dot-reinforced PVA composites as edible food packaging films and coatings with antimicrobial and UV-blocking properties. *ACS Omega*. 2022;7:29967-29983.
5. Ghorpade VS, Dias RJ, Mali KK, Mulla SI. Citric acid crosslinked carboxymethylcellulose-polyvinyl alcohol hydrogel films for extended release of water soluble basic drugs. *J Drug Deliv Sci Technol*. 2019;52:421-430.
6. Xue YL, Huang J, Lau CH, Cao B, Li P. Tailoring the molecular structure of crosslinked polymers for pervaporation desalination. *Nat Commun*. 2020;11:1461. doi:10.1038/s41467-020-15038-w
7. Dattola E, Parrotta EI, Scalise S, et al. Development of 3D PVA scaffolds for cardiac tissue engineering and cell screening applications. *RSC Adv*. 2019;9:4246-4257.
8. Khalaf M, Saeed AM, Ali AI, Kamoun EA, Fahmy A. Polyelectrolyte membranes based on phosphorylated-PVA/cellulose acetate for direct methanol fuel cell applications: synthesis, instrumental characterization, and performance testing. *Sci Rep*. 2023;13:1-14.
9. Sonker AK, Rathore K, Nagarale RK, Verma V. Crosslinking of polyvinyl alcohol (PVA) and effect of crosslinker shape (aliphatic and aromatic) thereof. *J Polym Environ*. 2018;26:1782-1794.
10. Chung Y. The effects of grafting and crosslinking aliphatic dicarboxylic acids onto polyurethane on the tensile stress and shape recovery. *Macromol Res*. 2018;26(9):1-6.
11. Nambiar AP, Pillai R, Vadikkeettil Y, Sanyal M, Shrivastav PS. Glutaraldehyde-crosslinked poly(vinyl alcohol)/halloysite composite films as adsorbent for methylene blue in water. *Mater Chem Phys*. 2022;291:126752.
12. Itou T, Kitai H, Shimazu A, Miyazaki T, Tashiro K. Clarification of cross-linkage structure in boric acid doped poly(vinyl alcohol) and its model compound as studied by an organized combination of x-ray single-crystal structure analysis, raman spectroscopy, and density functional theoretical calculation. *J Phys Chem B*. 2014;118:6032-6037.
13. Jhang JC, Lin JH, Chen YS. Genipin-crosslinking polyvinyl alcohol hollow braids degradable tissue engineering scaffolds: manufacturing techniques and property evaluations. *J Appl Polym Sci*. 2021;138:1-9.
14. Mansur AAP, Rodrigues MA, Capanema NSV, Carvalho SM, Gomes DA, Mansur HS. Functionalized bioadhesion-enhanced carboxymethyl cellulose/polyvinyl alcohol hybrid hydrogels for chronic wound dressing applications. *RSC Adv*. 2023;13:13156-13168.
15. Zhang Y, Fang K, Wang W, Niu H. A water-soluble epoxy-based green crosslinking system for stabilizing PVA nanofibers. *Molecules*. 2022;27:4177. doi:10.3390/molecules27134177

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

16. Chen J, Han S, Huang M, Li J, Zhou M, He J. Green cross-linked nanofibers membrane based on CS/PVA combined with polybasic organic acid for tympanic membrane repair. *Int J Polym Mater Polym Biomater*. 2022;71:291-301.
17. Pangon A, Saesoo S, Saengkrit N, Ruktanonchai U. Multicarboxylic acids as environment-friendly solvents and in situ crosslinkers for chitosan/PVA nanofibers with tunable physicochemical properties and biocompatibility. *Carbohydr Polym*. 2016;138:156-165.
18. Gautam L, Warkar SG, Ahmad SI, Kant R, Jain M. A review on carboxylic acid cross-linked polyvinyl alcohol: properties and applications. *Polym Eng Sci*. 2022;62:225-246.
19. Xu S, Shen L, Li C, Wang Y. Properties and pervaporation performance of poly(vinyl alcohol) membranes crosslinked with various dianhydrides. *J Appl Polym Sci*. 2018;135:15-19.
20. Suganthi S, Mohanapriya S, Raj V. Tunable physicochemical and bactericidal activity of multicarboxylic-acids-crosslinked polyvinyl alcohol membrane for food packaging applications. *ChemistrySelect*. 2018;3(40):11167-11176.
21. Sonker AK, Tiwari N, Nagarale RK, Verma V. Synergistic effect of cellulose nanowhiskers reinforcement and dicarboxylic acids crosslinking towards polyvinyl alcohol properties. *J Polym Sci A Polym Chem*. 2016;54:2515-2525.
22. Le TH, Vuu MND, Hoang QM, Nguyen NT, Mai PT. Pervaporation dehydration of ethanol-water mixture using cross-linked poly(vinyl alcohol) membranes. *Sci Technol Dev J*. 2016;19:97-106.
23. Dlamini DS, Wang J, Mishra AK, Mamba BB, Hoek EMV. Effect of cross-linking agent chemistry and coating conditions on physical, chemical, and separation properties of PVA-Psf composite membranes. *Sep Sci Technol*. 2014;49:22-29.
24. Gohil JM, Bhattacharya A, Ray P. Studies on the cross-linking of poly (vinyl alcohol). *J Polym Res*. 2006;13:161-169.
25. Xie Z, Hoang M, Duong T, Gray S. Sol-gel derived poly (vinyl alcohol)/silica hybrid membrane–synthesis and characterisation. *J Membr Sci*. 2010;383:1-16.
26. Kudoh Y, Kojima T, Abe M, Oota M, Yamamoto T. Proton conducting membranes consisting of poly(vinyl alcohol) and poly(styrene sulfonic acid): crosslinking of poly(vinyl alcohol) with and without succinic acid. *Solid State Ion*. 2013;253:189-194.
27. Mahmoud A, Fahmy A, Naser A, Saied MA. Novel sulfonated poly (vinyl alcohol)/carboxy methyl cellulose/acrylamide-based hybrid polyelectrolyte membranes. *Sci Rep*. 2022;12:1-15.
28. Pandit AH, Mazumdar N, Imtiyaz K, Rizvi MMA, Ahmad S. Periodate-modified gum Arabic cross-linked PVA hydrogels: a promising approach toward Photoprotection and sustained delivery of folic acid. *ACS Omega*. 2019;4:16026-16036.
29. Larrañeta E, Imízcoz M, Toh JX, et al. Synthesis and characterization of lignin hydrogels for potential applications as drug eluting antimicrobial coatings for medical materials. *ACS Sustain Chem Eng*. 2018;6:9037-9046.
30. Gupta P, Purwar R. Influence of cross-linkers on the properties of cotton grafted poly (acrylamide-co-acrylic acid) hydrogel composite: swelling and drug release kinetics. *Iran Polym J*. 2021;30:381-391.
31. Madhurima V, Sudheendran K, Raju KCJ. Ab initio and dielectric studies of succinic acid and maleic acid in 1,4-dioxane. *Mol Simul*. 2006;32:331-337.
32. Liu T, Peng X, Chen Y, Zhang J, Jiao C, Wang H. Solid-phase esterification between poly(vinyl alcohol) and malonic acid and its function in toughening hydrogels. *Polym Chem*. 2020;11:4787-4797.
33. Truong YB, Choi J, Mardel J, et al. Functional cross-linked electrospun polyvinyl alcohol membranes and their potential applications. *Macromol Mater Eng*. 2017;302:1-9.
34. Apelblat A, Manzurola E. Solubility of oxalic, malonic, succinic, adipic, maleic, malic, citric, and tartaric acids in water from 278.15 to 338.15 K. *J Chem Thermodyn*. 1987;19:317-320.
35. Suganthi S, Vignesh S, Kalyana J, Vairamuthu S. Fabrication of PVA polymer films with improved antibacterial activity by fine tuning via organic acids for food packaging applications. *Appl Water Sci*. 2020;10:1-11.
36. Hasan AMA, Keshawy M, Abdel-Raouf MES. Atomic force microscopy investigation of smart superabsorbent hydrogels based on carboxymethyl guar gum: surface topography and swelling properties. *Mater Chem Phys*. 2022;278:125521.
37. Kumar A, Bhatt R, Katare R. Atomic force microscopy enabled roughness analysis of nanostructured poly (diaminonaphthalene) doped poly (vinyl alcohol) conducting polymer thin films. *Micron*. 2016;90:12-17.
38. do Nascimento FC, de Aguiar LCV, Costa LAT, et al. Formulation and characterization of crosslinked polyvinyl alcohol (PVA) membranes: effects of the crosslinking agents. *Polym Bull*. 2021;78:917-929.
39. Hatakeyema T, Yamauchi A, Tsukuba Y, Hatakeyema H, Tsukuba Y. Studies on bound water in poly(vinyl alcohol). Hydrogel by DSC and FT-NMR. *Eur Polym J*. 1984;20:61-64.
40. Dou Y, Zhang B, He M, Yin G, Cui Y, Savina IN. Keratin/polyvinyl alcohol blend films cross-linked by dialdehyde starch and their potential application for drug release. *Polymers*. 2015;7(3):580-591.
41. Park JS, Park JW, Ruckenstein E. Thermal and dynamic mechanical analysis of PVA/MC blend hydrogels. *Polymer*. 2001;42:4271-4280.
42. Benavente R, Mijangos C, Peren JM, Krumova M, Lo D. Effect of crosslinking on the mechanical and thermal properties of poly (vinyl alcohol). *Polymer*. 2000;41:9265-9272.

How to cite this article: Gautam L, Warkar SG, Jain M. Physicochemical evaluation of polyvinyl alcohol films crosslinked with saturated and unsaturated dicarboxylic acids: A comparative study. *Polym Eng Sci*. 2024;1-13. doi:10.1002/pen.26806

15482634, 0, Downloaded from https://4spepublications.onlinelibrary.wiley.com/doi/10.1002/pen.26806 by Delhi Technological University, Wiley Online Library on [11/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ONLINE FIRST June 24, 2024

Computer Assisted Methods in Engineering and Science
***, doi: 10.24423/cames.2024.722

This article belongs to the *Special Issue on AI-Based Future Intelligent Networks and Communications Security* edited by Dr. S. Kumar, Dr. G. Mapp, Dr. A. Bansal and Dr. K. Cengiz

Predicting Wealth Score from Remote Sensing Satellite Images and Household Survey Data Using Deep Learning

Shashank SHEKHAR[1], Pratibha SINGH[2], Rashmi MISHRA[3], Sunil KUMAR[4]

[1)] *Edunet Foundation, Gurugram, India*

[2)] *Department of Computer Science & Engineering, Krishna Engineering College, Ghaziabad, India*

[3)] *Delhi Technological University, India*

[4)] *Institute for Communication Systems, University of Surrey, United Kingdom*

[*] *Corresponding Author e-mail: sk4sunilkumar@gmail.com*

The most exigent call of the United Nations' 17 sustainable goals is to end poverty everywhere by 2030. Unlike in the past, when poverty was measured based on data collected through ground-level surveys, the new technology adopted by many developing and developed countries is to estimate the poverty index using remote sensing satellite images with the help of machine learning techniques. Our approach demonstrates the prediction of cluster wealth score and establishes the relationship between wealth score obtained from Demographic and Health Survey (DHS) data and remote sensing satellite images of India by calculating Pearson's correlation coefficient (r^2). The implementation results have been analyzed in four phases. Phase 1 comprises four regression models (RMs): Ridge, RANSAC, Lasso, and k-nearest neighbor for feature extraction from a pre-trained convolutional neural network model using daylight & nightlight images. Here, the Lasso RM outperforms the others and is best suited for predicting the wealth score. Phase 2 categorizes daylight images with DHS data, where the Lasso RM efficiently generates the cluster wealth score. Phase 3 focuses on images of specific regions of Delhi, Tamil Nadu, Maharashtra and Telangana, using the Lasso RM, as it emerged as the best predictor of cluster wealth score in the previous two phases. Phase 4 compares the results attained through our proposed model with existing results.

Keywords: convolutional neural network, Demographic and Health Survey Data, Inception V3, transfer learning.

Copyright © The Author(s).
Published by IPPT PAN. This work is licensed under the Creative Commons Attribution License CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The word poverty comes from the old Latin word *pauper*, which means *poor* in English. According to the World Bank, poverty can be regarded as the deprivation of well-being, including low income and the inability to acquire the basic goods and services essential for survival with dignity. The eradication of extreme poverty in all its forms is the major goal of the world's leading organizations, such as the United Nations (UN), the World Bank, etc. The mission of eliminating poverty is rooted in all the analytical, operational and united work in more than 145 client countries. Despite significant strides in reducing extreme poverty, the rate of reduced poverty remains high in low-income countries and those affected by conflict and political instability. Progress has been marked in reducing poverty, which is the first of the world's Sustainable Development Goals (SDGs) as mentioned by the UN [1].

As India is a developing nation, poverty is declining in the country with each passing day. According to Brookings reports, it is estimated that "India may become the third-largest economy by 2030 and poverty will fall below 10 percent by 2025". The reported data also showed that India will be out of the list of top 10 countries in terms of extreme poverty by 2030 [8]. The most common method used in India to estimate poverty is typically based on income or consumption levels, and if either parameter falls below the given minimum level, the person is considered to be in debt or living below the poverty line.

The government and its numerous bodies run several programs and implement different policies to alleviate poverty. Economists and intellectuals have different opinions regarding the estimation of poverty, which has led to divergent views and confusion in the country. In India, the earlier method used to estimate poverty involved collecting survey data on the ground, which was not always accurate, was a costly method of data collection, and consumed a lot of time [2]. The most novel approach used nowadays to estimate the poverty index is using techniques such as machine learning (ML) and deep learning (DL) combined with remote sensing satellite data. This method effectively overcomes the shortcomings of traditional data collection and is able to project poverty at a large scale in developing countries. The scheme has been successfully implemented and has yielded better outcomes in several African countries like Nigeria, Uganda, Tanzania and Rwanda.

In this paper, we have effectively used a pre-trained convolutional neural network (CNN) model on daylight satellite images for predicting the wealth score of different regions. We have also observed that there is a strong correlation between predictions obtained through daylight images with nightlight images and daylight images labelled using household survey data. The contributions of this paper are as follows:

1. Exploring the role of recent developments in ML for predicting the wealth score of regions in any country. This is prerequisite for utilizing limited available resources to reduce poverty in the most affected regions. It will help in attaining the most significant SDG of the UN, emphasizing the importance of addressing this problem and exploring its solutions.
2. Analyzing CNN's performance along with four different RMs: Ridge, Lasso (least absolute shrinkage and selection operator), RANSAC (random sample consensus) and KNN (k-nearest neighbor) for identifying the correlation of wealth score obtained from household survey data with high-resolution satellite images (daylight and nightlight) of various regions of India.
3. Learning robust features from daylight images categorized using nightlight images or survey data as available, using a pre-trained Inception V3 model. Transfer learning (TL) solves the problem of availability of insufficient labeled data and reduces the computational efforts of training a network from scratch.
4. As we know, collecting household data through ground-level surveys is very expensive and time-consuming, the use of TL has been encouraged by combining a CNN with different RMs to predict wealth score utilizing satellite images, thus reducing the role of survey data.
5. Forecasting wealth score will help government policy makers and NGOs to effectively allocate resources to regions with deficiencies and improve their economic conditions.

2. Literature survey

The scarcity of data concerning livelihood, social, health and environmental issues produces unreliable information regarding economic livelihoods in the developing world [3]. Due to the low number of surveys, developing countries have very little data to make significant decisions. In contrast, developed countries have a lot of data to measure poverty. In the list of SDGs released by the UN, poverty is the first of the 17 objectives to be eradicated by 2030 [8]. Traditionally, the major source for measuring poverty was collecting data from ground-level surveys such as household consumption and wealth surveys. However, these methods were inaccurate, tedious, and costly. The living standard measurement survey (LSMS) was initially used in the context of ground-level surveys. Household expenditure was measured using LSMS approach. The limitation of LSMS surveys was that they did not provide comprehensive information needed to accurately calculate the poverty in a particular region [4, 5].

The use of mobile data and digital footprint has also emerged as a method to calculate poverty. Frequent data on societal, economic, and livelihood conditions is required to track SDGs and target poverty-related information. New technologies such as ML and data science have now evolved as new measures to predict poverty. Researchers have begun to develop these techniques to calculate poverty [6]. One of the most highly structured, easily available, and inexpensive data sources is remote sensing data, such as satellite imagery. Using satellite imagery makes it easier to minimize the data gap between actual values and predicted values. For this purpose, a CNN is trained to estimate poverty levels using satellite imagery [7, 8]. With ML techniques, it is also possible to establish the relationship of predicting poverty over several years. These techniques learn visual features from satellite imagery which indicate the poverty levels in a region.

Several experiments also suggest that poverty prediction from different and multiple parameters gives more accurate results compared to using a single parameter [9]. Object identification, image segmentation and labelling of satellite images help to identify visual features to estimate poverty. The United States Government's Landsat 7 program was used to acquire publicly available satellite imagery. The model is pre-trained on one of the datasets, and further, this model can be trained to predict nightlight intensity corresponding to daylight satellite imagery. Daytime satellite imagery can be used to estimate local per capita outcomes [4]. Features learned from one model can be used to train other models. This technique is known as transfer learning (TL). A Google Geocoding API is employed to predict the location of a place's center as well as latitude and longitude. Multi-task learning involves learning multiple tasks simultaneously by analyzing the similarities and differences among the tasks [10, 11]. The suitable input image size for a model is 1920×1920 pixels. For predicting wealth, nightlight images serve as a good proxy. Socio-economic indicators can be best extracted with the help of raw satellite images. There is a need for more frequent data that can learn features from the TL approach to monitor the range of issues covered by 17 goals and 169 targets [12]. The TL approach not only predicts the nightlight intensity corresponding to daylight satellite images but also helps in mapping the poverty level.

The ultimate goal of TL approach is to transfer the knowledge from daylight satellite imagery to the target problem for predicting nightlight intensity. TL is highly scalable and does not require human annotation. However, nightlight intensity faces difficulties in distinguishing between densely populated, poor and sparsely populated areas, making it unsuitable for estimating per capita consumption [13]. The multitask model can be used for predicting poverty and at the same time evaluating the similarities and differences among the tasks. The multitask fully convolutional model comprises convolutional layer,

ReLU activation layer, batch normalization layer, pooling layer, and a dropout layer [14].

The model is able to establish the relationship between specific graphic structures and growth factors. These features reflect the relationship between ground-truth values and predicted values. Pearson's correlation coefficient (PCC) calculation is used to assess model accuracy. Differences in the relationship between household data and remote sensed data are evident at the community level [8]. The prediction of household-level wealth in a particular area can be visualized with the help of a classification tree. To check whether the TL model improves with nightlights, 210 trials of 10-fold cross-validation were conducted. Another model to predict poverty is the Wikipedia embedding model, trained using the mean squared error loss function and Adam optimizer. The Wikipedia learning model outperforms using nightlights only [31]. The poor performance of the model is due to the limitations of the nightlight baseline. To forecast poverty predictions across years and determine their accuracy a recurrent neural network can be integrated on top of a CNN.

To revolutionize social science, remote sensing data presents new potential, especially with the night-time light dataset that provides night-time luminosity from 1993 to 2013 [17]. It evaluates the night-time light data against the rural electrification data from the 2011 Census of India. The results suggest that these night-time satellite images are accurate for estimating rural electrification. The accuracy is improved with better GIS maps, basic geoprocessing tools, and night-time luminosity aggregation [6]. The nighttime satellite images and census data were also used to correlate light intensity from images with state-wise poverty, population, GDP, and forest cover. The ARIMA model is used for forecasting the census data. The output is measured in terms of PCC with r^2 value of four states of India namely [19]:

1. Uttar Pradesh (0.68),
2. Himachal Pradesh (0.68),
3. Jharkhand (0.37) and
4. Gujarat (0.25).

The summary of literature survey is shown in Table 1. As seen in the literature survey of different countries, we conclude that wealth score of any area can be predicted with the help of satellite images such as daylight and nightlight images. There are several other methods also to predict the poverty of any area such as LSMS Surveys, Object Identification, UN Govt. Landsat's Program, Wikipedia Embedding Model, and Tracking of Mobile Call Records. Other technologies involved the use of ARIMA model. But the best among all that we have seen in the literature survey is the use of daylight and nightlight images that accurately predicts the wealth score or poverty of any area.

TABLE 1. Description of literature survey.

Title	Location	Datasets	Algorithm	Result (adjusted r^2)	Parameter		
Nischal *et al.* [6]	India	1) DMSP-OLS satellite images from the National Geophysical Data Center (NOAA), 2) State-wise GDP and population data for India from 2000–2012 from the Ministry of Statistics and Programme Implementation	Multivariate RM, The ARIMA (Auto Regressive Integrated Moving Average) Model used for Future Census Forecasts	Uttar Pradesh (0.68), Himachal Pradesh (0.68), Jharkhand (0.37), Gujarat (0.25)	Coefficient of determination		
Sheehan *et al.* [12]	Malawi, Nigeria, Tanzania, Uganda, Ghana	1) Demographic and Health Surveys (DHS) data, 2) Afrobarometer (Round 6): a survey across 36 countries in Africa, accessing various aspects of welfare, education and infrastructure quality, 3) Geolocated Wikipedia articles, 4) Visible Infrared Imaging Radiometer Suite (VIIRS) nightlight data	Nightlight-Only, Wikipedia embedding Model, Multi-modal Model	Location	Nightlight only	Wikipedia embedding	Multi-modal
				Ghana	0.41	0.47	0.76
				Malawi	0.24	0.49	0.64
				Tanzania	0.6	0.64	0.71
				Nigeria	0.3	0.52	0.7
				Uganda	0.53	0.57	0.76
Pandey *et al.* [13]	India	1) 2011 census data (Uttar Pradesh Province), 2) Google Static Maps API	Multi-task fully CNN	0.82	Coefficient of determination		
Duan *et al.* [14]	New York City (USA)	1) Felony Dataset – contains several incidents, with the properties of coordinates, offense type and time, 2) 311 datasets	CNN, Support Vector Machine (SVM), Random Forest (RF), SpatioTemporal Crime Network (STCN)	SVM (0.74), RF (0.82), STCN (0.92), SFCNN (0.82)	F1 Score, AUC ROC		
Ruchir [16]	Malawi, Nigeria, Tanzania, Uganda	1) Living Standard Measurement Survey (LSMS) by the World Bank, 2) Landsat 7 imagery (US Govt.), 3) VIIRS Nightlight Data	ResNet, Ridge RM	Malawi (0.55), Nigeria (0.68), Tanzania (0.57), Uganda (0.69)	Coefficient of determination		

Table 1. [Cont.].

Title	Location	Datasets	Algorithm	Result (adjusted r^2)	Parameter
Eugenie *et al.* [2]	India	1) India Lights Project's API to download village-level data, 2) Nighttime Lights Dataset (by NOAA) Linear	Linear regression	0.63	Coefficient of determination
Neal *et al.* [3]	Rwanda	1) Demographic & Health Surveys (DHS) – the primary source for population and health statistics, 2) Expenditure data from the World Bank's LSMS surveys	CNN transfer learning	0.75	Coefficient of determination
Xie *et al.* [19]	Uganda	1) ImageNet, 2) DHS Survey	CNN transfer learning	71% accuracy	Cross-validation scheme test
Watmough *et al.* [20]	Sauri (western Kenya)	1) Remote sensing data/ satellite data, 2) Daylight images	Classification	0.62	Coefficient of determination
Perez *et al.* [21]	Uganda	1) Landsat 7 (African Continent), 2) DHS data (Africa)	Wasserstein generative adversarial networks (GANs)	0.57	Coefficient of determination
Piagessi *et al.* [22]	Santiago, Los Angeles, Chicago, Boston, Houston, Philadelphia	1) Household income data (Santiago, Chile), 2) Satellite data from DigitalGlobe web platform, 3) Satellite images (Google Static Maps API)	ResNet 50, VGGF, VGGF+Nightlights	Santiago (0.52) Los Angeles (0.56) Chicago (0.36) Boston (0.37) Houston (0.327) Philadelphia (0.46)	Coefficient of determination
Steele *et al.* [23]	Bangladesh	1) DHS data 2011, 2) National household Surveys, 3) Call detail record (Grameenphone)	Hierarchical Bayesian geo-statistical methods	0.78	Coefficient of determination
Catamuro *et al.* [24]	Kenya	Road quality measurement and corresponding satellite imagery	CNN, transfer learning, ResNet, AlexNet, VGG-11	0.79 0.35	Regression standard training-test split, Harder held-out regression problem

Table 1. [Cont.].

Title	Location	Datasets	Algorithm	Result (adjusted r^2)	Parameter		
Ajami *et al.* [9]	Bangalore	Very high resolution (VHR) satellite images and Set of geographic information system (GIS) layer	VGG	0.75	Coefficient of determination		
Guie *et al.* [15]	China	1) DMSP-OLS satellite images from National Geophysical Data Center (NGDC)	Gaussian process with radial basis function kernel (GPRBFK), partial least squares regression for generalized linear models (PLSRGLM)	GP (84.6%), PLS (82.8%), RF (83.3%)	Accuracy		
Subash *et al.* [17]	India	1) India Lights API, 2) Satellite Images from the Defense Meteorological Satellite Program (DMSP), 3) Per capita GDP	Artificial neural network, ML algorithms	Root Mean Square Error ML (nightlights), ML (per capita GDP)	RSME 7.38 8.29	r^2 0.59 0.56	– –
Suraj *et al.* [25]	India	1) Census 2011, 2) Daylight Satellite Images (Google Static Maps API) (2017), 3) Nightlight data (Defense Meteorological Satellite Program's Operational Linescan System-DMSP-OLS)	VGG-CNN-S	0.65	Coefficient of determination		

3. Proposed method

The Phase 1 model, presented in Figs. 1 and 2, initiates with Demographic and Health Survey (DHS) data, i.e., household data (containing records for everyone) and geographical data (DHS data linked with health and infrastructure). By processing these two files, cluster and household data are generated, which contain wealth score, cluster wealth, latitude, and longitude for different locations. The raw survey data is then used to produce the more refined cluster and household data, containing wealth score, nightlight intensity, number of households, household ID, and cluster ID. The extracted cluster data is managed to store files that hold the values of each cluster size and the values of nightlight. Daylight images are downloaded from the signed URL of the HERE Maps API, sized at 400×400 pixels using generated image coordinates. Later, the size is reduced to 299×299 pixels, and the files generated contain the downloaded dataset locations. Similarly, from the DMSP-OLS, the nightlight images are downloaded at size $43\,201 \times 18\,801$ pixels.

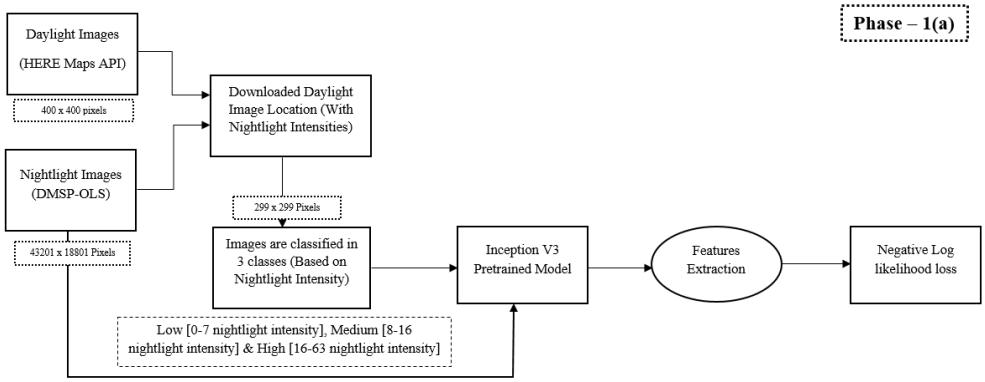


Fig. 1. Phase 1(a) for the training model with TL.

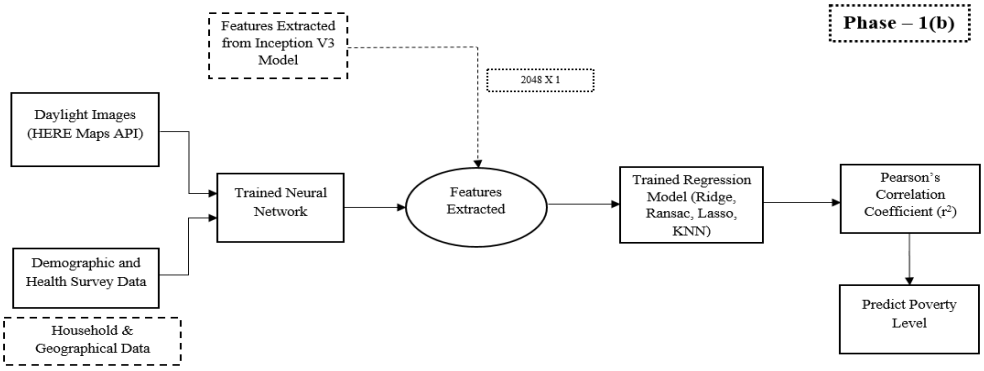


Fig. 2. Phase 1(b) for the regression model and predicting poverty.

The downloaded daylight images are then divided into three classes: low, medium, and high. With the help of the pre-trained Inception V3 model, we train our CNN that establishes the relationship between training, testing loss and accuracy. The features of satellite images are now extracted with the help of the trained model along with latitude and longitude values. Regression models such as Lasso, Ridge, RANSAC and KNN are trained using features generated from the trained CNN model and DHS data. We can use this DHS data to plot a single-panel regression value graph between model prediction and actual assets value for all four RMs. In the end, confusion matrices with normalization and without normalization are plotted for the trained model.

Similarly, in Phase 2 depicted in Fig. 3, the model initiates with daylight images only to predict poverty. These daylight images are classified and trained with the help of a pre-trained network, and different regression models are then applied to the daylight images only to predict poverty using PCC (r^2).

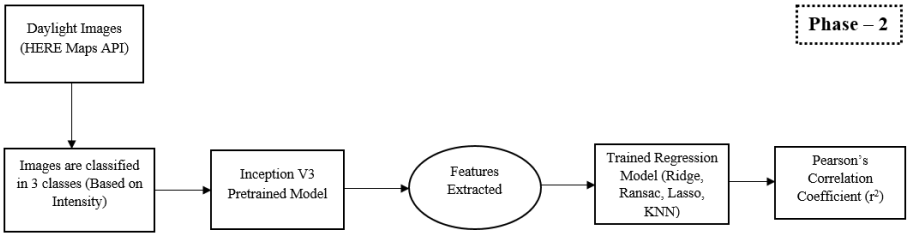


FIG. 3. Phase 2 for the training, regression and prediction of poverty.

The Phase 3 model proposes a method for the prediction of poverty, where daylight images and nightlight images of Delhi/ NCR, Tamil Nadu, Maharashtra, and Telangana regions are considered. The locations of images are identified with the help of latitude and longitude values generated. These images are then divided into three categories: low, medium, and high, based on their nightlight intensity. These images are then processed through the Inception V3 pre-trained model for feature extraction. To predict poverty in all three regions, their images will be trained with the Lasso RM using PCC (r^2).

3.1. Regression models

3.1.1. The Ridge regression model (Ridge RM). It is a technique for handling multiple regression data that suffer from multicollinearity issues. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will give estimates that are more reliable. The cost function for ridge regression is shown in Eq. (1):

$$\sum_{i=1}^{M}(y_i-\widehat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j*x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2, \tag{1}$$

where x – the predictor variable, y – the response variable, $\widehat{y}_i$ – the predicted value of y for the i-th instance, M – the number of instances in the dataset, λ – the regularization penalty, p – the number of features, and w – the number of coefficients.

In ridge regression, the cost function is transformed by penalty term that is equal to the sum of squared coefficients multiplied by a tuning parameter λ. It puts a limitation on the coefficients in such a way that a penalty is imposed on the coefficient function if it takes large values. With ridge regression, the value of coefficients is minimized and the complexity of the model is reduced. If the value of λ becomes 0, the cost function of ridge regression reduces to the linear model cost function [27].

3.1.2. The Lasso regression model (Lasso RM). Lasso RM is very similar to Ridge regression. While Ridge regression penalizes the sum of squared coefficients (L2 penalty), Lasso penalizes the absolute value sum (L1 penalty) for non-zero coefficients. The loss function for Lasso regression is shown in Eq. (2):

$$\sum_{i=1}^{M}(y_i-\widehat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j*x_{ij}\right)^2+\sum_{j=0}^{p}|w_j|. \tag{2}$$

Equation (2) differs from Eq. (1) in that Lasso regression uses the magnitude of coefficients in place of the square of coefficients. This means that many coefficients may be disregarded. Lasso RM helps in feature selection as well as reducing overfitting [27].

3.1.3. The RANSAC regression model (RANSAC RM). It is a learning technique used to estimate model parameters for observed data through random sampling. If a dataset contains both inliers and outliers, RANSAC RM finds the optimal fitting result by using a voting scheme [28].

3.1.4. The KNN regression model (KNN RM). It is used for both classification and regression problems. To predict the value of new points of data, the concept of feature similarity is used. The new value of the data point is assigned based on the closest point in the training data [29].

3.1.5. Pearson's correlation coefficient (PCC). PCC is a statistical measure used to compute the linear correlation between two variables: X and Y. The

value for PCC ranges between 1 and −1, where +1 is total positive linear correlation, 0 is no linear correlation and −1 is total negative linear correlation. PCC is calculated as the covariance of the two variables X and Y divided by their standard deviations. The formula for PCC is given in Eq. (3)

$$\rho_{X,Y} = \frac{E\left[XY\right] - E\left[X\right]E\left[Y\right]}{\sqrt{E\left[X^2\right] - \left[E\left[X\right]\right]^2}\sqrt{E\left[Y^2\right] - \left[E\left[Y\right]\right]^2}}, \tag{3}$$

where $E[X]$ – the mean of X scores, $E[Y]$ – the mean of Y scores, and $E[XY]$ – the mean of the product of paired score.

3.2. Inception V3 architecture

Inception V3 is a CNN trained on an ImageNet dataset [26] to enhance image analysis and object detection. This ImageNet Dataset is comprised of more than 14 000 000 images. The size of each input image in the dataset is 299×299 pixels. Inception V3 model is a well-designed convolutional module that produces discriminative features and reduces the number of parameters. The inception module is composed of several convolutional layers, pooling layers, and fully connected layers. This model consists of a total of 32 convolutional layers of different sizes, 5 (3 avg and 2 max) pooling layers, and 5 fully connected layers with different numbers of neurons, including 2048, 2048 and 210. The base network of Inception V3 is formed with the combination of convolutional and pooling layers [18].

The output of the Inception V3 network consists of 210 classes. The layer-wise classification of the Inception V3 model is given above in Table 2.

Table 2. Inception V3 architecture [18].

Model	nCLs	Filter sizes	nPLs	Filter sizes	nFCs	No. of neurons
Inception V3	15	1×1				
	7	3×3				
	3	1×7	3 (avg)	3×3	3	2048, 2048, 210
	3	7×1	2 (max)	3×3		
	1	1×3				
	3	3×1				

3.3. Dataset description

3.3.1. Daylight images. The daylight satellite images used in the paper are downloaded from the HERE Maps API, where the dimension of each image

is 400×400 pixels, and there are 60 000 images. The obtained latitudes and longitudes help to download the images from the HERE Maps API [10].

3.3.2. Nightlight intensities. The nightlight images are used to obtain the nightlight intensities. These nightlight images are taken from DMSP-OLS data source accessible from 1990s to the present. Each image is of size $43\,201 \times 16\,801$ pixels [4].

3.3.3. Cluster wealth score. The wealth index is calculated as an integral part of the inspection of goods. An inquiry is made into questions about the ownership of common items such as bicycles, televisions, and materials used for house construction. These parameters are normalized for each country and therefore we do not need further generalizations [5].

4. Implementation and results

We tested the same concept with a fixed learning rate (LR) value of 0.001 and 0.0001. We predicted the accuracy of the model on a pre-trained model (true and false) with these two LR values. Ridge RM was tested for predicting the accuracy and cluster value. With an LR value of 0.001 and pre-trained set to true, the best test accuracy achieved was 78.54%.

System configuration used for implementing the proposed approach is shown in Table 3.

Table 3. System configuration.

Python	3.8.1
Jupyter notebook	6.0.3
R	3.6.2
Processor	I7-9750H CPU (9th Gen)
RAM	16 GB DDR4
GPU	4GB GTX 1650

4.1. Phase 1: daylight + nightlight images

Table 4 presents the results of training the model with the help of daylight and nightlight Images. The model is trained with LR values set to 0.02, 0.42, and 0.89, and the momentum and epoch are kept at 0.9 and 20, respectively. The Inception V3 retrain model is used to train our model, and RMs are trained to predict the wealth score. As shown in Table 4, accuracy and loss analysis

Table 4. Accuracy and loss analysis for different values of learning rate.

Daylight + nightlight images						
Model	Accuracy and loss					
	Batch size = 64					
Pre-trained	Momentum/Epoch	Learning rate	Training accuracy	Training loss	Test accuracy	Test loss
True	0.9/20	0.02	99.1	0.35	80.1	1.7
		0.42	94.7	0.51	75.3	1.16
		0.89	88.7	1.16	76.9	1.11
	0.9/30	0.001	99.9	0.35	78.5	92.8
		0.0001	65.7	1.11	64.2	77.4
False	0.9/30	0.001	55.1	1.28	58.7	87.1
		0.0001	49.5	1.38	50.5	96.3

is done on different values of LR, i.e., 0.02, 0.42, and 0.89, revealing that the training test accuracies are higher when the LR value is low.

The value of test loss decreases with decreasing LR in some cases, However, this is not consistent across all LRs. We infer from the table that training and testing accuracy are higher when the value of LR is low. The results for LR values of 0.001 and 0.0001 for both pre-trained = true and pre-trained = false were also included [2].

LR is a hyper-parameter that controls the weights of the neural network with respect to the loss gradient. It defines how quickly the neural network updates the concept it has learned. Adaptive LR allows the training algorithm to monitor the performance of the model and automatically adjust the LR for the best performance. Neural networks with the help of adaptive LR can usually outperform those fixed values of LR.

4.1.1. Learning rate: 0.02. As shown in Fig. 4, the training loss gradually degraded from the 1st to the 7th iteration, then slightly curved down from the 6th to the 8th iteration before reaching a saturation state after the 9th iteration. The test loss shows frequent fluctuations until the 7th iteration, after which it starts to increase gradually and reaches saturation after the 13th iteration. On the other hand, the training accuracy gradually increased from the very first iteration until the 7th iteration and after the 7th iteration, it remained constant due to the fixed LR.

The test accuracy shows repeated fluctuations from different iterations and stabilizes from the 7th iteration without further deflection. Figure 5 presents the training of the model with the help of four different RMs: Ridge, RANSAC,

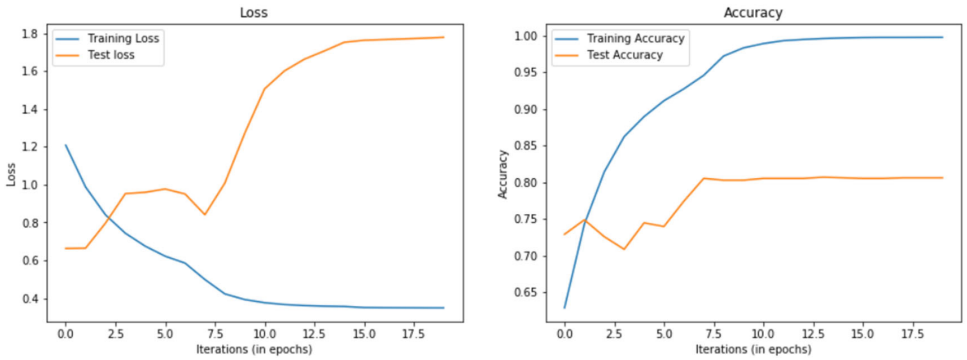


FIG. 4. Iteration vs. loss and accuracy graph.

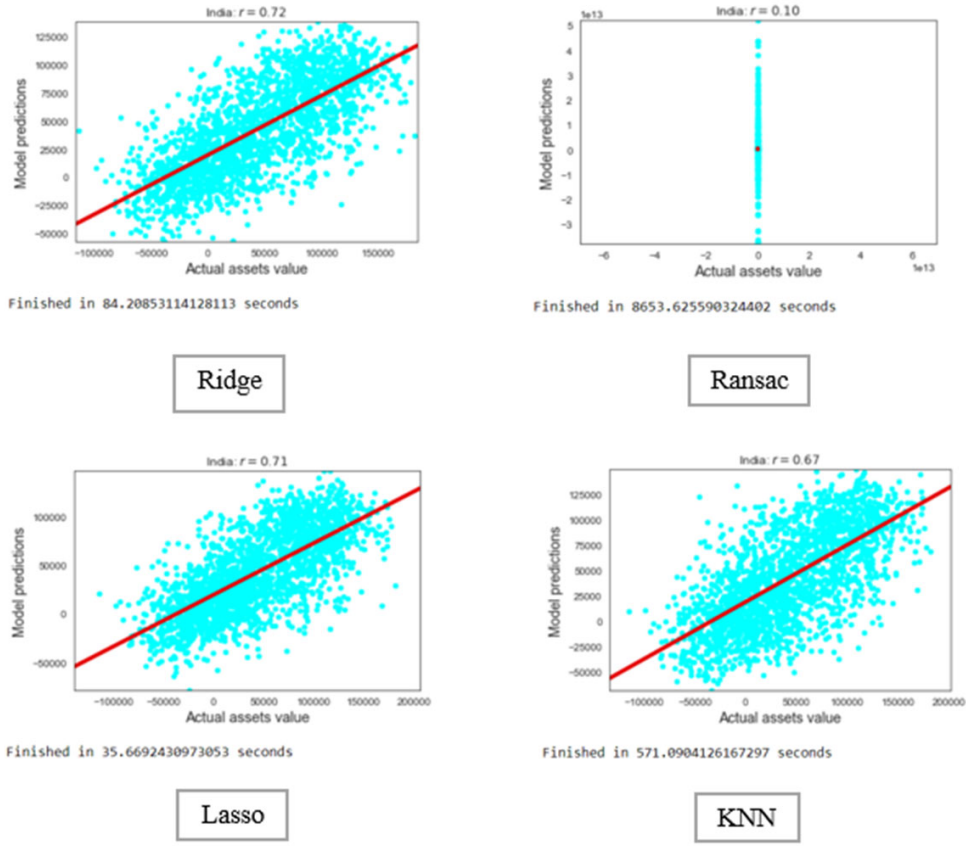


FIG. 5. Comparison of clusters with different RMs for predicting wealth score.

Lasso and KNN, where the Lasso RM requires significantly less time to train the model, achieving an r^2 value of 0.71.

For all three class labels (low, medium and high), confusion matrices with and without normalization are shown above in Figs. 6a and 6b.

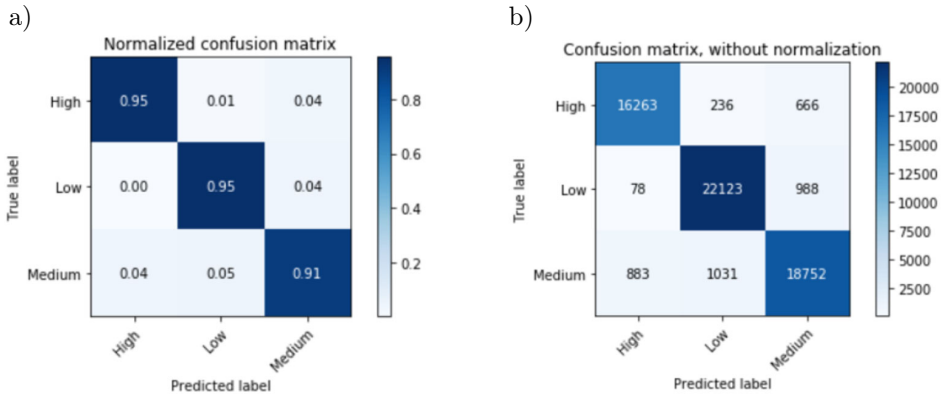


FIG. 6. Confusion matrices with standardization (a) and without standardization (b).

4.1.2. Learning rate: 0.42. As presented in Fig. 7, the training loss o starts from a very high value and then decreases after experiencing a sharp downward deflection at the 6th iteration, reaching saturation after the 11th iteration. Contrarily, the test loss starts with a value of 1 and without showing very much deflection it starts increasing very slowly after the 12th iteration. At the same time, the training accuracy starts from a very low value and exponentially increases from the very first iteration until the last iteration.

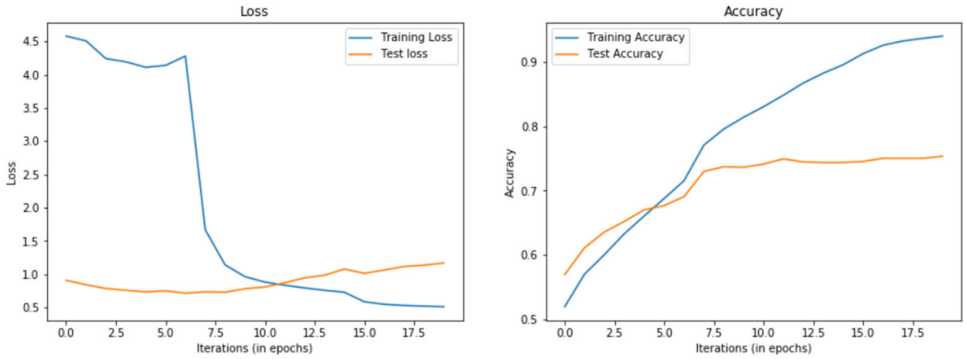


FIG. 7. Iteration vs. loss and accuracy graph.

Conversely, the test accuracy gradually increases at the beginning but shows some deflection due to the slowly decreasing LR and fixed momentum between the 5th and 10th iterations, after which the test accuracy saturates from the 11th iteration onward.

The results depicted below in Fig. 8 show the performance differences between various RM with the LR set at 0.42. The PCC (r^2) for Ridge, Lasso, RANSAC and KNN RMs is 0.62, 0.65, −0.01 and 0.61, respectively. A negative PCC here signifies that the straight-line model through the data explains $0.01^2 = 0.0001 = 0.99\%$ of the data variance, and the (–) signifies that the slope of the graph is negative.

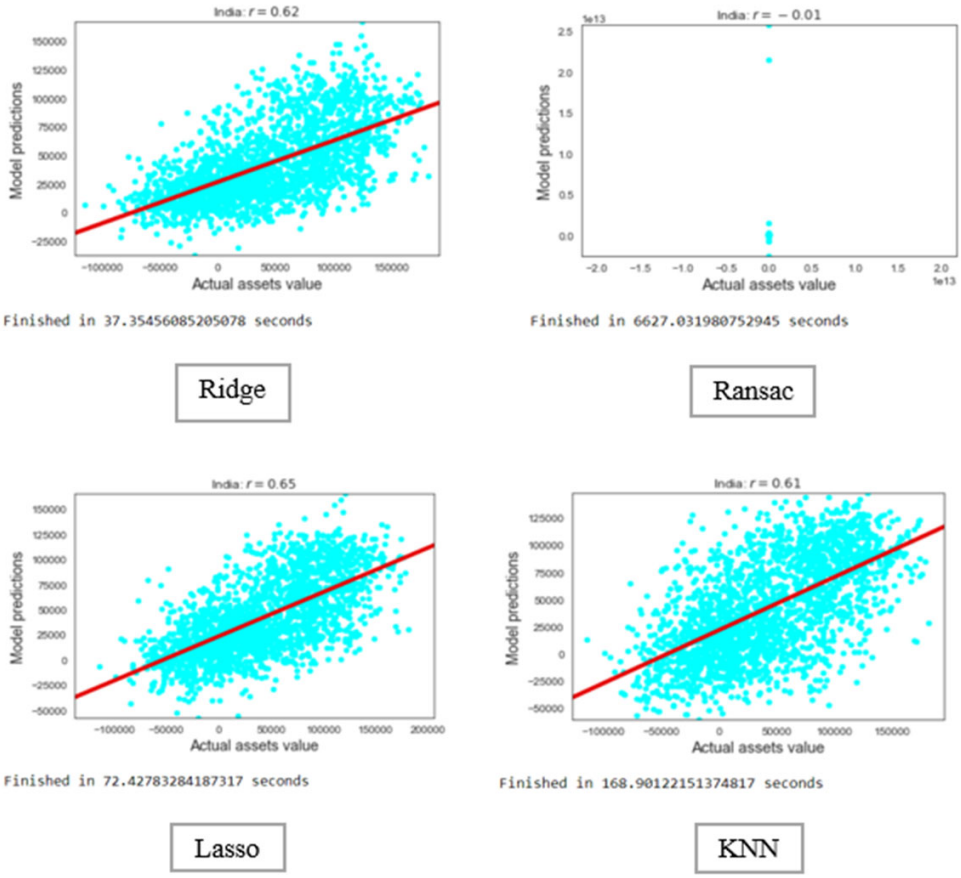


FIG. 8. Comparison of clusters with different RMs for predicting wealth score.

Here, the Lasso RM outperforms the other RMs, achieving the highest accuracy of the model with an r^2 value of 0.65.

Confusion matrices for three output classes (low, medium, and high) are shown in Figs. 9a and 9b.

4.1.3. Learning rate: 0.89. As shown in Fig. 10, the training loss deflects sharply and frequently initially, followed by sharp decrease from the 6th iteration, and reaching a saturation state after the 8th iteration onward.

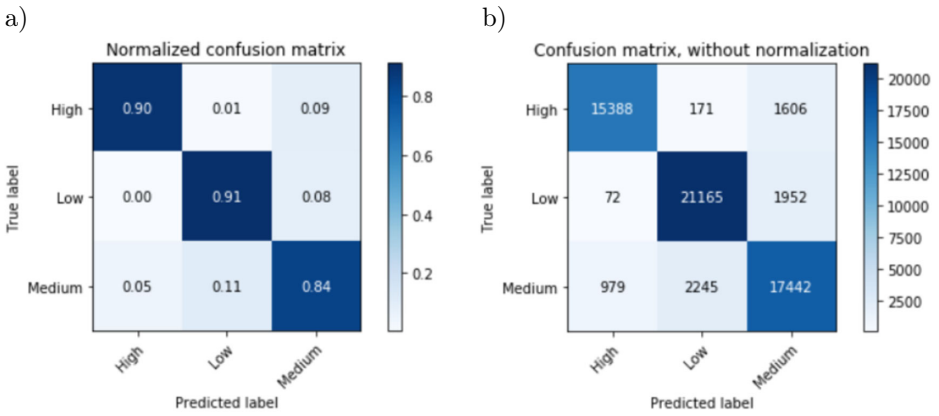


FIG. 9. Confusion matrices with standardization (a) and without standardization (b).

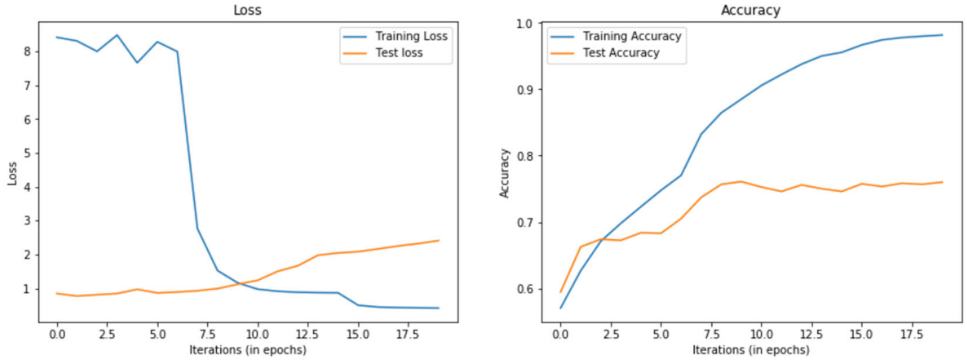


FIG. 10. Iteration vs. loss and accuracy graph.

This is due to the fast-paced training of the model with high LR of 0.89. Conversely, the test loss starts increasingly very smoothly from the very first iteration, and by the 13th iteration, it reaches a saturation state with some deflection. Additionally, the training accuracy exponentially increases from the very first iteration to the last iteration. However, the test accuracy shows repeated fluctuations with each iteration. It keeps on changing its values regardless of the fixed LR and momentum.

Figure 11 shows performance difference in four RMs with an LR value of 0.89. The PCC (r^2) is 0.64, 0.66, −0.02, and 0.61 for Ridge, Lasso, RANSAC, and KNN RMs, respectively. Among these, Lasso RM with LR 0.89 proved to be the best RM that predicted the best accuracy on the dataset provided. RANSAC RM again shows a negative correlation coefficient, indicating that an increase in the LR value leads to a decrease in the PCC value of the provided dataset.

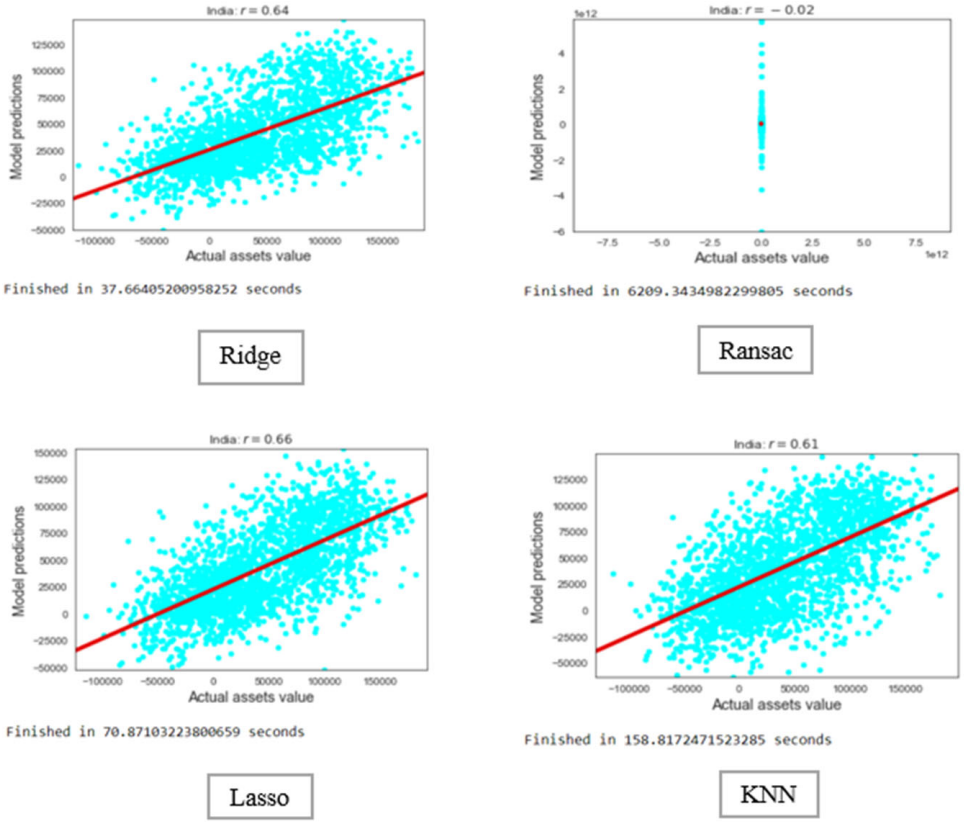


Fig. 11. Comparison of clusters with different RMs for predicting wealth score.

Confusion matrices with and without normalization are once again displayed for LR value of 0.89 and low, medium and high classes in Figs. 12a and 12b.

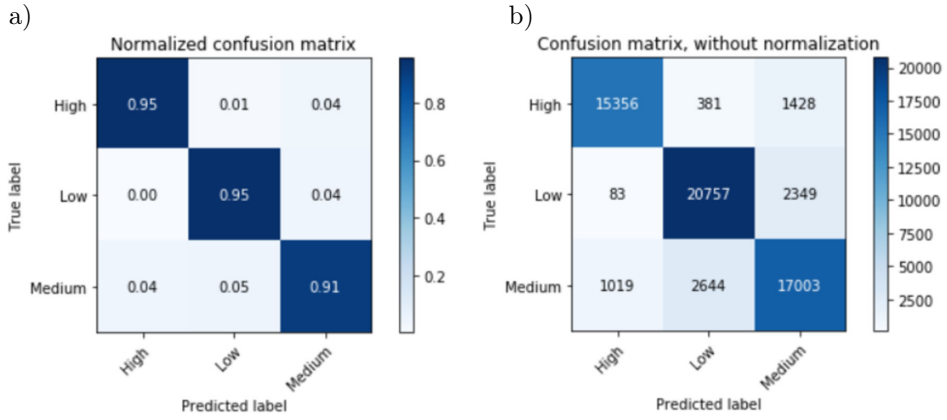


Fig. 12. Confusion matrices with standardization (a) and without standardization (b).

4.2. Phase 2: daylight images only

In this phase, the model is trained with the help of daylight images only. The locations of the images are identified with the help of latitude and longitude obtained from the Household and Geographical Survey Data. Table 5 shows the analysis of accuracy and loss with different LR values. It is observed that the training accuracy and test accuracy are highest for lower values. It is also observed that train loss decreases when we decrease LR. The momentum and number of epochs taken here are set to 0.9 and 20, respectively. The results are obtained using the daylight images only.

TABLE 5. Accuracy and loss analysis with various values of LR.

Daylight + nightlight images						
Model	Accuracy and loss					
	Batch size = 64					
Pre-trained	Region	Momentum/Epoch	Learning rate	Training accuracy	Training loss	Test accuracy
True	Delhi/NCR	0.9/20	0.02	0.958	0.294	0.777
	Tamil Nadu	0.9/20	0.02	0.967	0.471	0.841
	Maharashtra	0.9/20	0.02	0.877	0.605	0.866

4.2.1. Learning rate: 0.02. Figure 13 shows the graph plotting the training and testing loss/accuracy vs. iteration. The training loss degrades from the very 1st iteration and saturates after the 10th iteration.

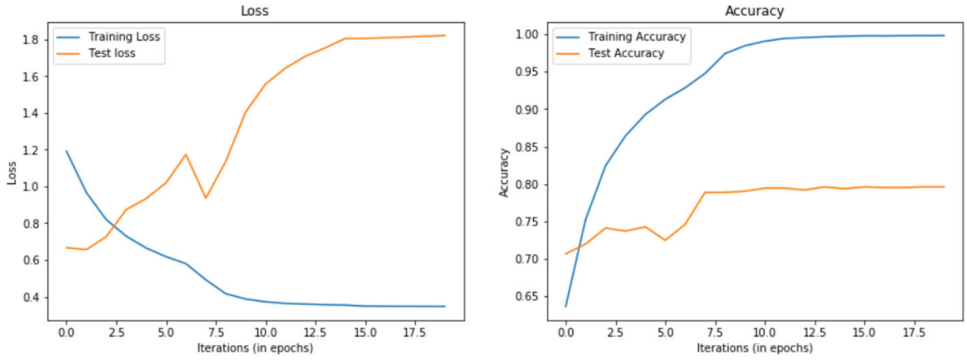


FIG. 13. Iteration vs. loss and accuracy graph.

The testing loss turns sharply at the 6th iteration and again at the 7th iteration. Next, it exponentially increases after the 7th iteration due to the LR decrement by a factor of 10. Meanwhile, the training accuracy exponentially

increases from the very first iteration and saturates as LR decreases gradually. The test accuracy shows frequent deflections before the 7th iteration and then stabilizes after the 7th iteration.

Figure 14 shows the performance comparison of four different RMs-Ridge, Lasso, RANSAC and KNN, where the Lasso RM outperforms all the other RMs with a PCC value of 0.72, completing the process in 35.42 sec. This indicates that with LR of 0.02, Lasso RM gives a better output on daylight images.

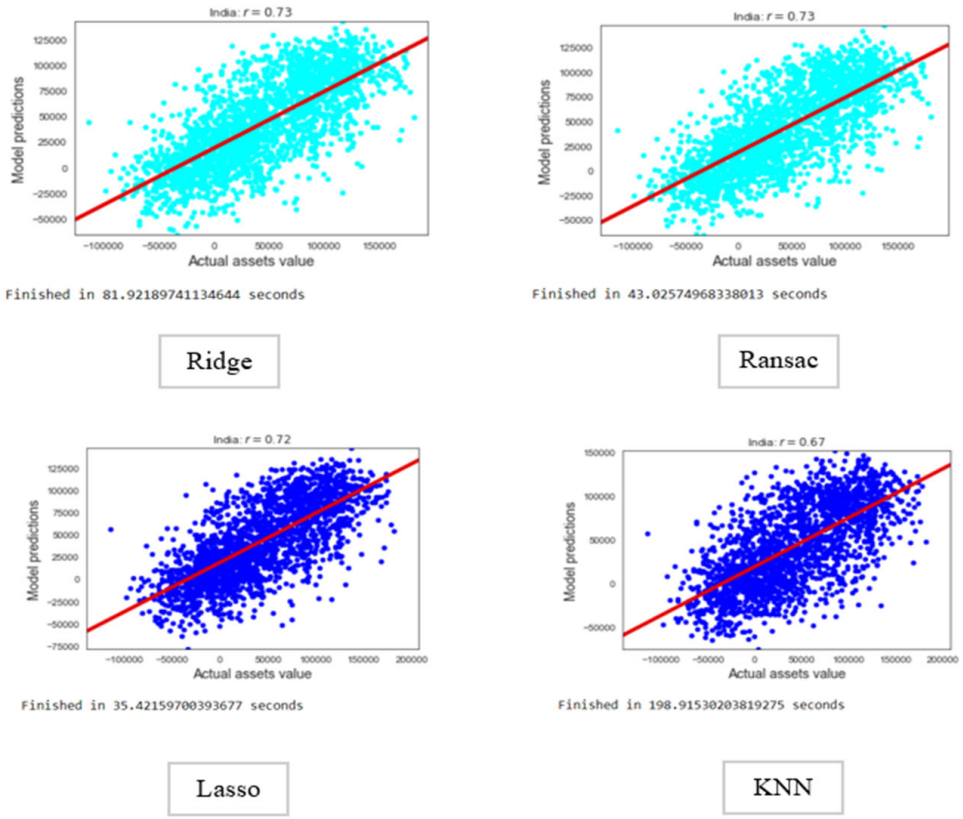


Fig. 14. Comparison of clusters with different RMs for predicting wealth score.

Confusion matrices with normalization and without normalization for LR = 0.02 are given in Figs. 15a and 15b.

4.2.2. Learning rate: 0.42. The training loss starts with a high value from the very 1st iteration and shows frequent deflections until the 6th iteration, then gradually decreases and saturates after the 8th iteration due to changes in LR. In contrast, the test loss starts with a very low value and maintains a single value throughout the iterations. Both training and testing accuracy exponen-

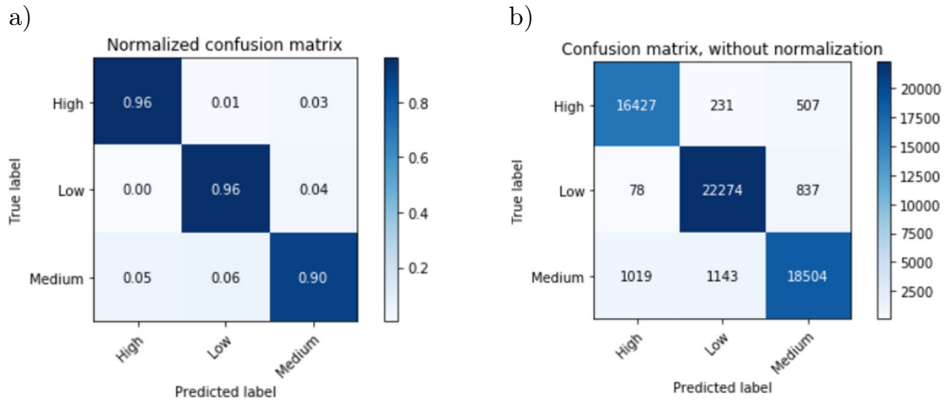


FIG. 15. Confusion matrices with standardization (a) and without standardization (b).

tially increase from the very 1st iteration and converge at the 7th iteration due to a single fixed value of LR, as shown in Fig. 16.

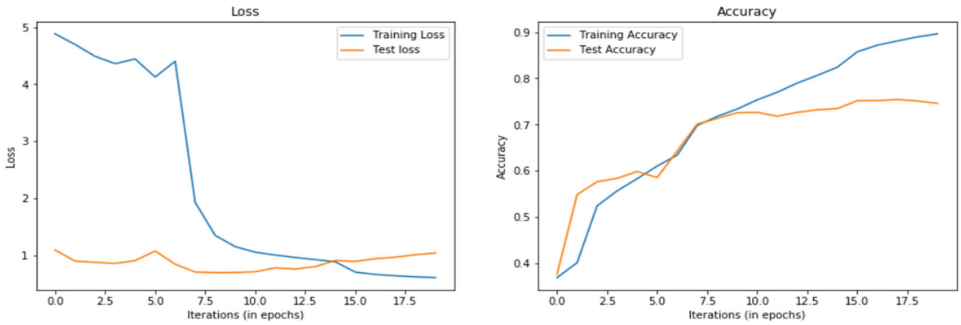


FIG. 16. Iteration vs. loss and accuracy graph.

As shown in Fig. 17, the four RMs are trained on daylight images with LR value = 0.42. Here, with this LR value, the RANSACRM outperforms all three models with an r^2 value of 0.62 and a total processing time of 161.7 sec, which is comparatively faster than the other three models (Ridge, Lasso and KNN).

Confusion matrices with normalization and without normalization for LR = 0.02 are given in Figs. 18a and 18b, respectively.

4.2.3. Learning rate: 0.89. The training loss, shown in Fig. 19, demonstrates a slight transformation at the 2nd iteration before exponentially falling after the 6th iteration and reaching saturation after the 8th iteration due to the higher LR value. Consecutively, the test loss remains constant throughout all iterations. Both training accuracy and testing accuracy start increasing simultaneously from the 1st iteration, and have the same value from the 3rd iteration to the 6th

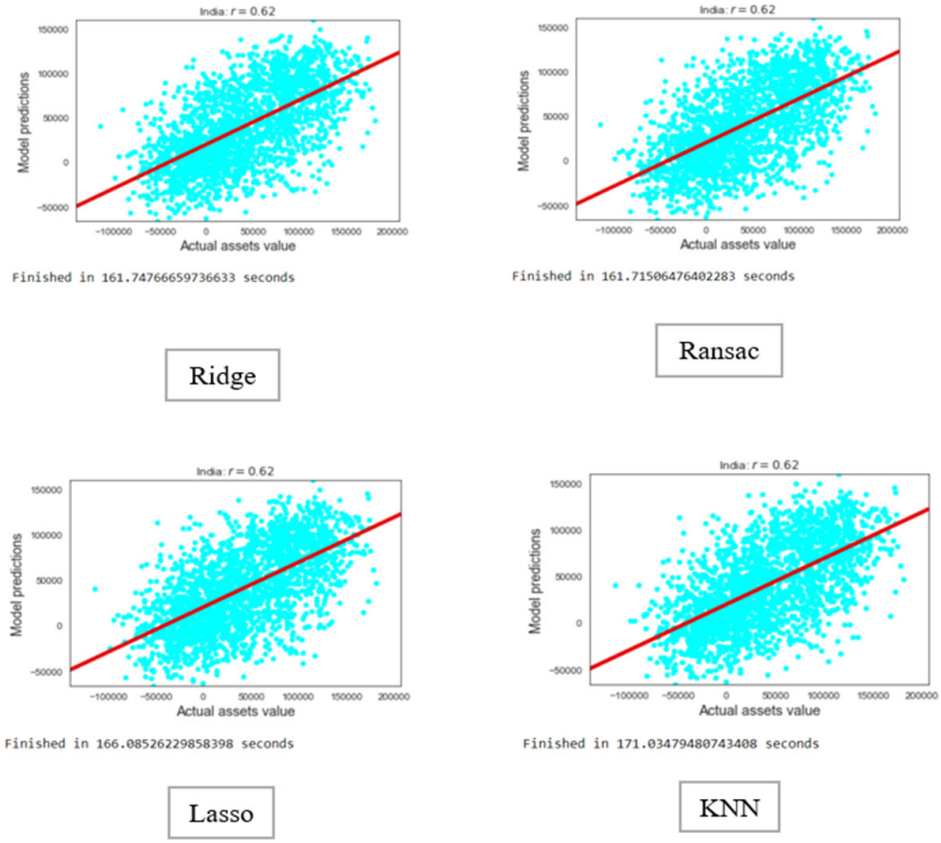


Fig. 17. Comparison of cluster with different RMs for predicting wealth score.

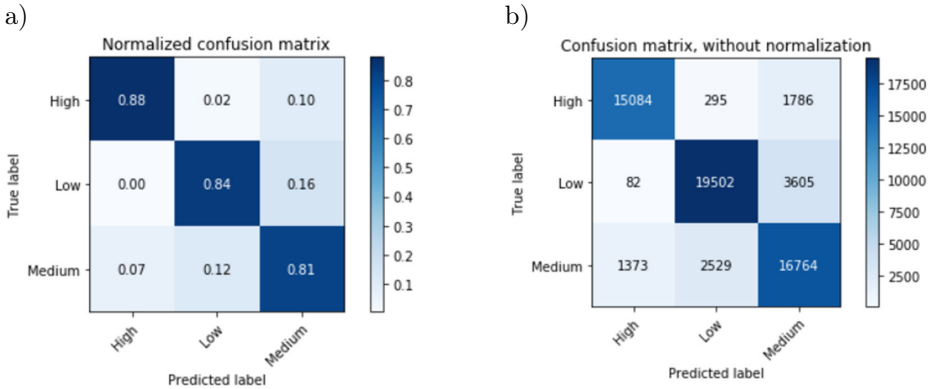


Fig. 18. Confusion matrices with standardization (a) and without standardization (b).

iteration, after which training accuracy increases rapidly while testing accuracy saturates from the 6th iteration.

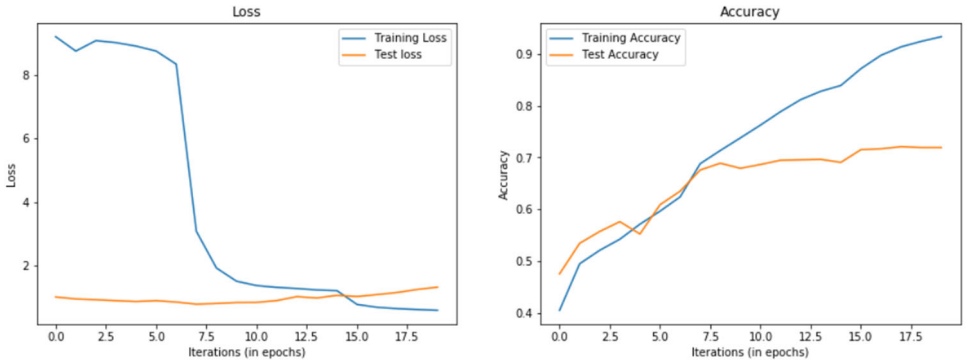


FIG. 19. Iteration vs. loss and accuracy graph.

The four regression models trained on daylight images are shown in Fig. 20, which clearly identifies that Lasso RM with an r^2 value = 0.63 is the best regression model for estimating PCC. Although the other TMs perform better at some points, Lasso RM predicts the coefficient with less processing time and fewer coefficient values.

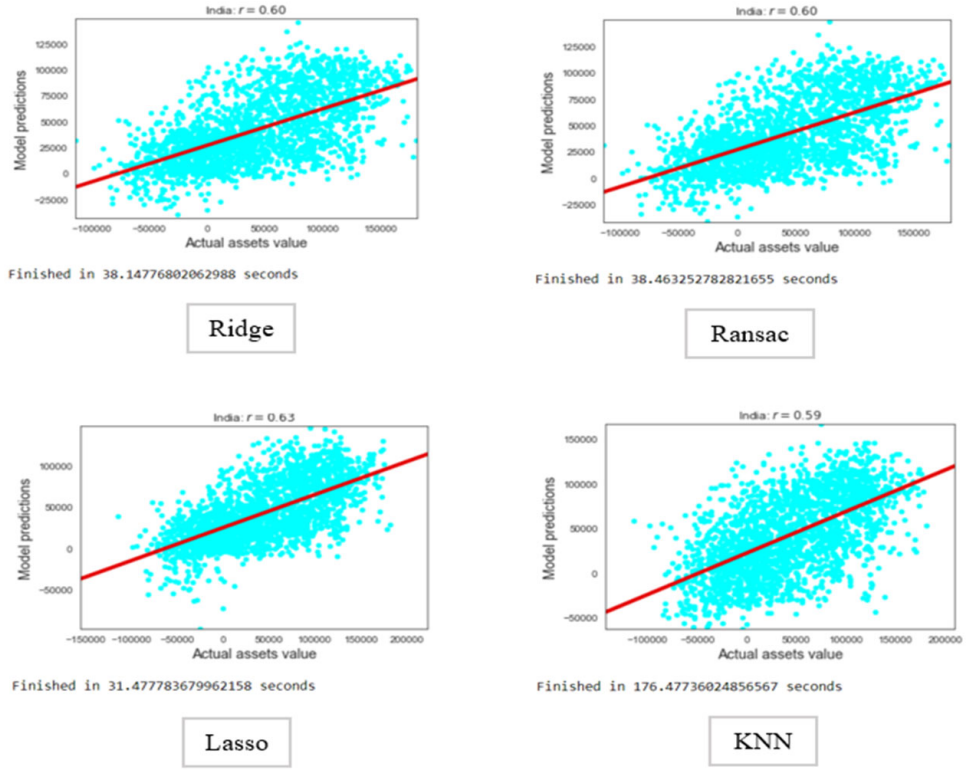


FIG. 20. Comparison of cluster with different RMs for predicting wealth score.

Confusion matrices with normalization and without normalization for LR = 0.02 are given in Figs. 21a and 21b, respectively.

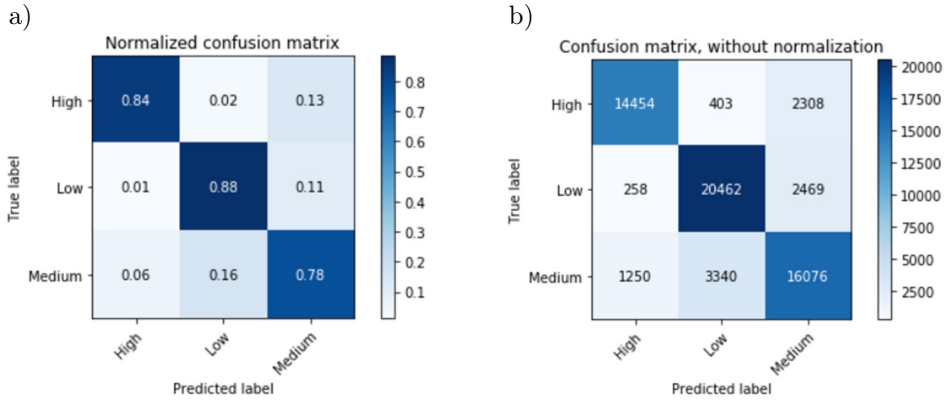


FIG. 21. Confusion matrices with standardization (a) and without standardization (b).

4.3. Phase 3

Table 6 shows the accuracy and loss analysis of training/testing conducted on daylight and nightlight images in the regions of Delhi/NCR, Tamil Nadu, and Maharashtra. The results are obtained using a single value of LR = 0.02, and the test accuracy for Delhi/NCR, Tamil Nadu, Maharashtra, and Telangana is 77.7, 84.1, 86.6 and 66.6, respectively.

TABLE 6. Accuracy and loss analysis for different regions.

Daylight + nightlight images							
Model	Accuracy and loss						
	Batch size = 64						
Pre-trained	Region	Momentum/ Epoch	Learning rate	Training accuracy	Training loss	Test accuracy	Test loss
True	Delhi/NCR	0.9/20	0.02	0.958	0.294	0.777	1.06
	Tamil Nadu	0.9/20	0.02	0.967	0.471	0.841	1.21
	Maharashtra	0.9/20	0.02	0.877	0.605	0.866	0.675
	Telangna	0.9/20	0.02	0.672	1.01	0.666	0.988

4.3.1. Delhi/NCR region (daylight + nightlight images). In this phase, we focus particularly on predicting poverty in the Delhi/NCR, Tamil Nadu, and Maharashtra regions of India. We filtered the images of these regions with the help of latitude and longitude values and trained the model for prediction with the help of the Lasso RM because it gave the best output with less time and

less coefficient value when trained with both daylight + nightlight images, as well as with daylight images alone. Sample images of the Delhi/NCR region are presented in Figs. 22 and 23.

FIG. 22. Sample image 1 [10].

FIG. 23. Sample image 2 [10].

The model is trained with the help of the Inception V3 pre-trained model with LR = 0.02, and momentum and number of epochs are kept at 0.9 and 20, respectively. The test accuracy of the model was determined to be 77.7%. Features extracted from the model were passed on to the and a comparison was made between PCC and DHS data.

Figure 24 presents the graph of training/testing loss and accuracy against iterations, where the training loss decreases exponentially until the 7th iteration before the saturation due to the constant LR value and parameters involved in the pre-training model, By contrast, the test loss shows frequent deflections until the 8th iteration and saturates after the 9th iteration. Training accuracy increases exponentially and saturates after the 3rd iteration, while test accuracy sharply increases until the 2nd iteration and saturates after the 7th iteration.

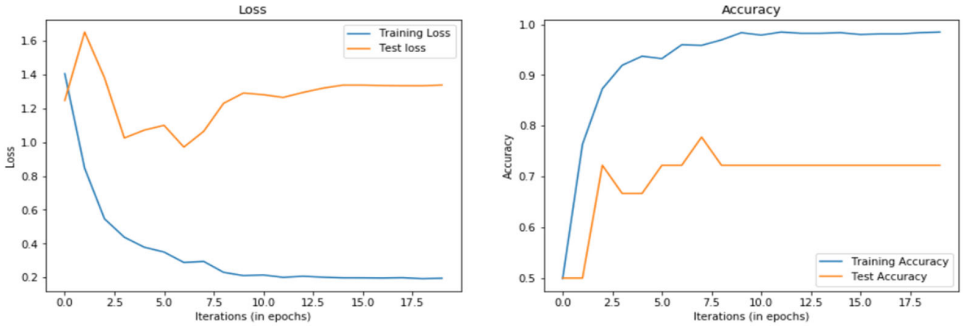


FIG. 24. Iteration vs. loss and accuracy graph.

The trained model is passed to the Lasso RM, which predicts PCC as 0.69 and takes 37.28 seconds. The prediction of the result is fairly accurate, considering

that cluster locations of DHS have eruption fused, upholding the invisibility of the survey in Fig. 25.

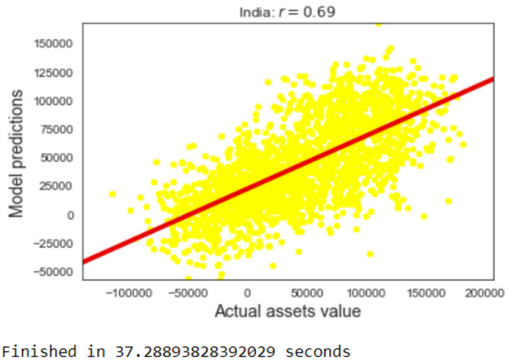


Fig. 25. Clusters for prediction of wealth score on daylight images.

Confusion matrices for the Delhi/NCR region (daylight + nightlight images) for low, medium and high classes are shown in Figs. 26a and 26b.

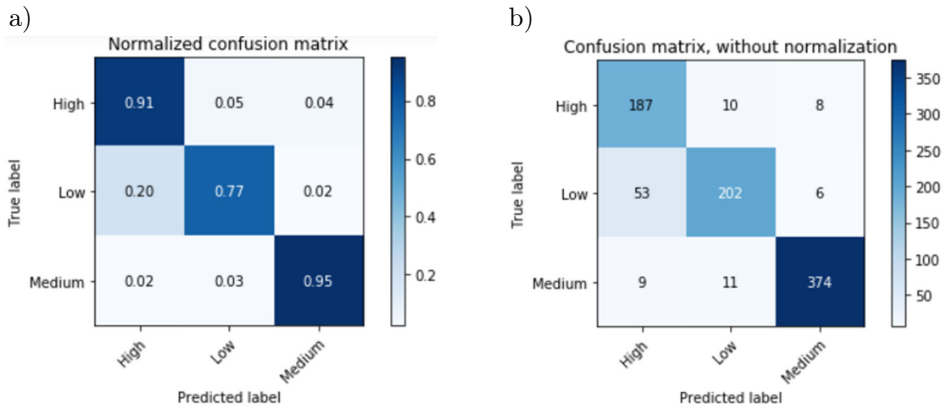


Fig. 26. Confusion matrices with standardization (a) and without standardization (b).

4.3.2. Tamil Nadu region (daylight + nightlight images). Sample images of the Tamil Nadu region are provided in Figs. 27 and 28.

As shown in Fig. 29, both test loss and training loss start with the same magnitude due to a fixed value of LR and momentum value, and the test loss shows frequent and sharp deflections until the 13th iteration before coming to a saturation state. The training loss continuously decreases with each iteration and saturates after the 12th iteration. The test accuracy with initial iteration shows very sharp deflection and then stabilizes during the iterations before finally saturating after the 13th iteration. The training accuracy gradually increases with each iteration.

FIG. 27. Sample image 1 [10].

FIG. 28. Sample image 2 [10].

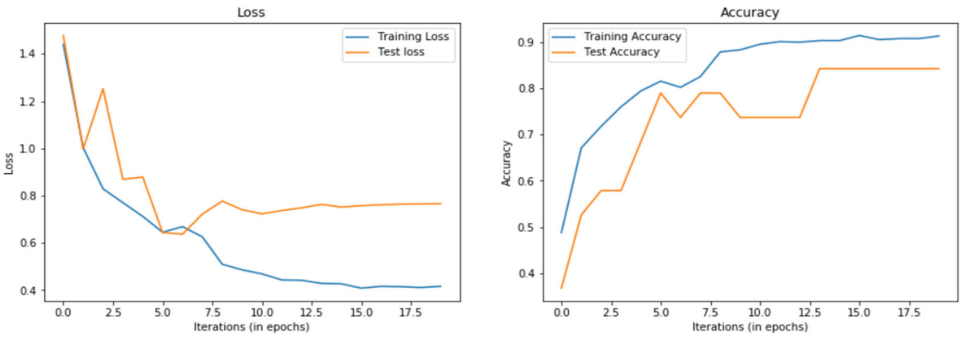
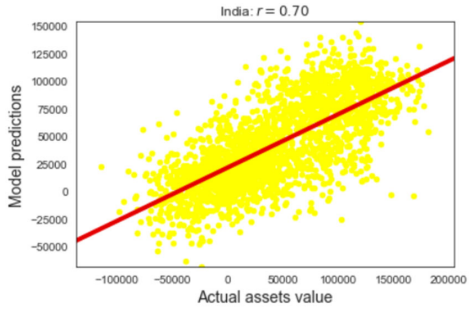


FIG. 29. Iteration vs. loss and accuracy graph.

The Lasso RM predicts PCC as 0.70 and takes 36.10 seconds. The prediction is quite accurate, considering that cluster locations of DHS have eruption fused maintaining the survey's invisibility, as shown in Fig. 30.




```
Finished in 36.107677936553955 seconds
```

FIG. 30. Clusters for prediction of wealth score on daylight images.

Confusion matrices for the Tamil Nadu region (daylight + nightlight images) for low, medium and high classes as shown in Figs. 31a and 31b.

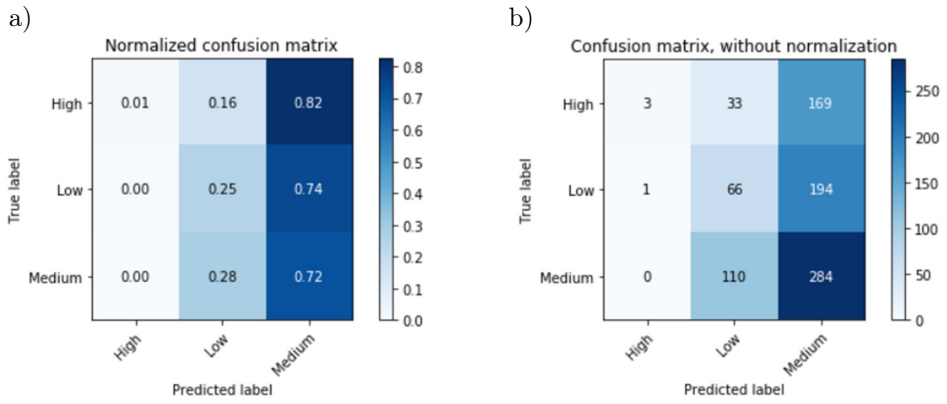


Fig. 31. Confusion matrices with standardization (a) and without standardization (b).

4.3.3. Maharashtra region (daylight + nightlight images). The third module deals with predicting the poverty index in the region of Maharashtra. Sample images of the Maharashtra region are given in Figs. 32 and 33.

Fig. 32. Sample image 1 [10].

Fig. 33. Sample image 2 [10].

As shown in Fig. 34, the test loss starts with a low value and exhibits sharp deflection from the 2nd to the 4th iteration before saturating after the 8th iteration. The training loss starts with a high magnitude and decreases gradually before coming to a saturation state after the 10th iteration. Test accuracy shows sharp variations before the 8th iteration and saturates after the 8th iteration. Training accuracy gradually increases with each iteration and shows no deflection after the 3rd iteration.

The model for the Maharashtra region is trained with Lasso regression, which predicts the value of r^2 as 0.69 and takes 76.10 seconds. The prediction is quite accurate considering that cluster locations of DHS have eruption fused that maintains the invisibility of the survey, as shown in Fig. 35.

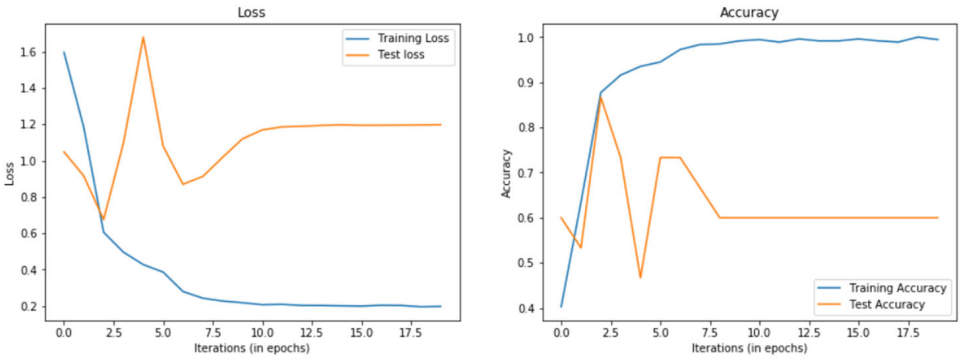
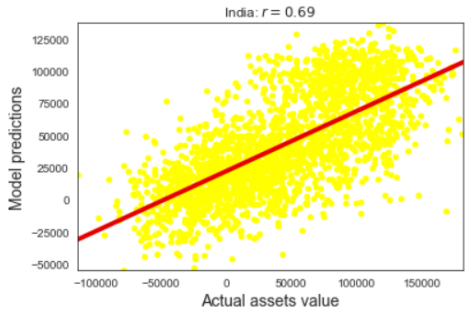


FIG. 34. The graph of iteration vs. loss and accuracy.




```
Finished in 76.10032796859741 seconds
```

FIG. 35. Clusters for prediction of wealth score on daylight images.

Confusion matrices for the Maharashtra region (daylight + nightlight images) for low, medium and high classes are shown in Figs. 36a and 36b.

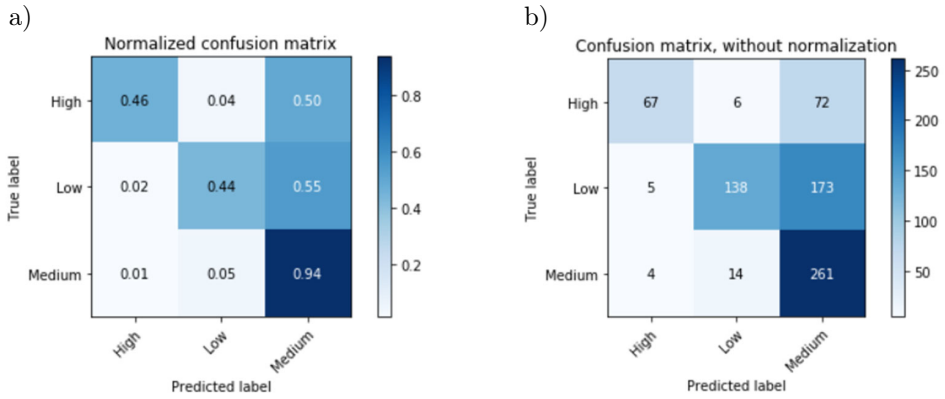


FIG. 36. Confusion matrices with standardization (a) and without standardization (b).

4.3.4. Telangana region (daylight + nightlight images). The sample images for the Telangana region are displayed in Figs. 37 and 38.

FIG. 37. Sample image 1 [10].

FIG. 38. Sample image 2 [10].

In Fig. 39, the training loss shows sharp deflections before the 5th iteration. After the 5th iteration, it starts increasing gradually and saturates after the 15th iteration. The test loss, on the other hand, stars with a very high value and degrades continuously until the 12th iteration before saturating after the 13th iteration. This is due to the very low value of LR and fixed value of momentum.

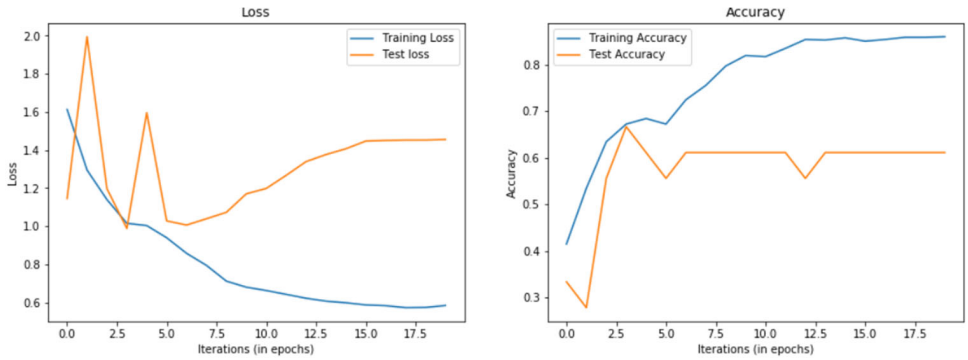


FIG. 39. Iteration vs. loss and accuracy graph.

Training accuracy gradually increases without significant variation in values, and saturates after the 13th iteration. Test accuracy, conversely, initially attains a very high value and decreases before the 5th, reaching a saturation status after the 6th iteration.

The model, trained with Lasso RM, yields an r^2 value of 0.68 and completes in 51.47 seconds. This suggests that the result obtained for the cluster score closely resembles the result obtained from the DHS survey, as shown in Fig. 40.

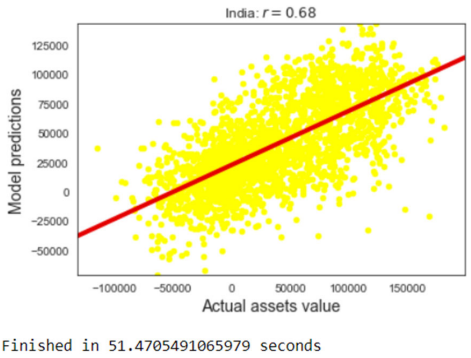


FIG. 40. Clusters for prediction of wealth score on daylight images.

Confusion matrices for the Telangana Region (daylight + nightlight images) for low, medium and high classes are shown in Figs. 41a and 41b.

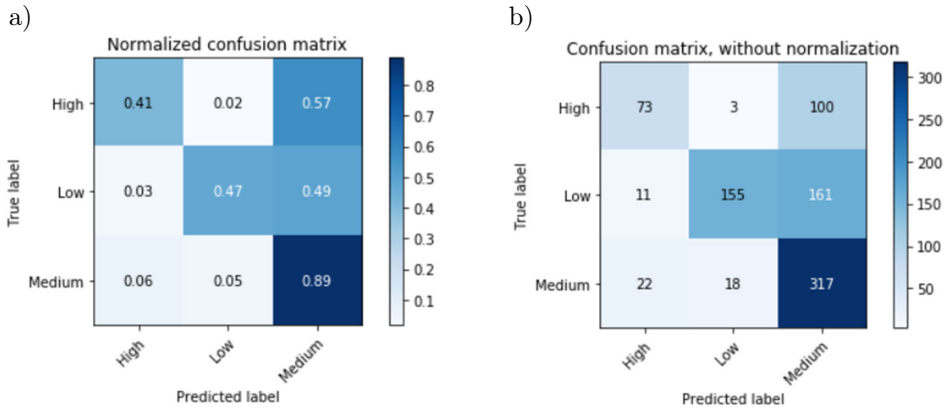


FIG. 41. Confusion matrices with standardization (a) and without standardization (b).

4.4. Phase 4: Comparison of results

As shown in Table 7, we have compared the results obtained from Phases 1 and 2 with respect to training and testing loss, accuracy and the coefficient of determination (R^2). Training accuracy remains the same in both phases and does not show much change in the values. The test loss also does not show much variance in the values. They remain close to 1 in both phases. We observed that the training loss decreases when using daylight images only. The coefficient of determination also improves when using daylight images only. The value of R^2 for daylight + nightlight images is 0.72 whereas for daylight images only, it is 0.73. This shows that using smaller data will improve the model's predictive performance.

Table 7. Comparison of results of different phases.

Comparison of results						
Momentum = 0.9, epoch = 20, batch size = 64						
Phases	Learning rate	Training accuracy	Training loss	Test accuracy	Test loss	R^2 (coefficient of determination)
1) Daylight + nightlight images	0.02	99.1	0.35	80.1	1.7	0.72
	0.42	94.7	0.51	75.3	1.16	
	0.89	88.7	1.16	76.9	1.11	
2) Daylight images only	0.02	99.8	0.34	79.6	1.8	0.73
	0.42	88.09	0.62	75.1	0.96	
	0.89	91.8	0.63	72.9	1.14	

The comparison of computation time is shown in Table 8, where the time taken by different RMs in Phases 1 and 2 is provided. We observed that the computation time for ridge RM does not show a significant difference for LR. However, a notable difference was observed in the case of the RANSAC RM, which required considerably more computation time when trained with both daylight and nightlight images compared to when trained with daylight images only required less time. Additionally, the computation time for the KNN RM shows a substantial difference when trained with LR of 0.02 and not much difference when trained with LR of 0.42 and 0.89. Notably, It was observed that Lasso RM takes shorter times compared to the other three RMs.

Table 8. Comparison of computation time.

Computation time [in seconds]					
Momentum = 0.9, epoch = 20, batch size = 64					
Phases	Learning rate	Ridge regression	Lasso regression	RANSAC regression	KNN regression
1) Daylight + nightlight images	0.02	84.2	35.6	8653.6	571
	0.42	37.3	72.4	6627	168.9
	0.89	37.6	70.8	6209.3	158.8
2) Daylight images only	0.02	81.9	35.4	43	198.9
	0.42	161.7	166	161.7	171
	0.89	38.1	31.4	38.4	176

As shown in Table 9, the results of our paper are compared with those of other papers in the same domain. The locations considered in other papers include India and several countries in Africa. The results obtained in our paper gave better results when compared with the results of several countries in Africa.

Table 9. Comparison of results of different research papers.

Parameter	Paper title	Location (value)
R^2 (coefficient of determination)	Multi-task DL for predicting poverty from satellite images [13]	India (0.82)
	Temporal poverty prediction in developing countries [16]	Africa: Malawi (0.55), Nigeria (0.68), Tanzania (0.57), Uganda (0.69)
	Combining satellite imagery and ML to predict poverty [3]	Rwanda (0.75)
	Socioecologically informed use of remote sensing data to predict the rural household poverty [20]	Sauri (Western Kenya) (0.62)
	Predicting city poverty using satellite imagery [22]	Chile: Santiago (0.52) USA: Los Angeles (0.56), Chicago (0.36), Boston (0.37), Houston (0.327), Philadelphia (0.46)
	Mapping poverty using mobile phones and satellite data [23]	Bangladesh (0.78)
	Identifying a slum degree of deprivation from VHR images using CNN [9]	India: Banglore (0.75)
	Predicting wealth score from remote sensing satellite images and household survey data using deep learning – current study	India (0.72) (daylight + nightlight) India (0.73) (daylight images) Delhi/NCR (0.69), Tamil Nadu (0.70), Maharashtra (0.69), Telangana (0.68)

Additionally, we provide a comprehensive comparison for both the entire area of India and four specific states – Delhi/NCR, Tamil Nadu, Maharashtra and Telangana.

5. Conclusion

We have exhaustively reviewed the literature to understand how ML techniques can help predict the socioeconomic parameter (poverty index) using remote sensing satellite images, household and geographical survey data. This review led us to analyze the use of satellite images and pre-trained models to predict poverty levels. We established a strong correlation between the results obtained from satellite images and DHS Data. In Phase 1, we considered both daylight and nightlight images. The poverty index was determined to be 0.72 using the Lasso RM. In Phase 2, we predicted the labels of the daylight images using survey data. Fine-tuning of the CNN improved the extraction of features, and the poverty index increased to 0.73. In both phases, Lasso RM outperformed

Ridge, RANSAC and KNN RMs. Considering the results obtained in Phases 1 and 2 using Lasso RM, we trained the model again in Phase 3 with images only from the Delhi/NCR, Tamil Nadu, Maharashtra and Telangana regions of India. The calculated wealth scores for these regions were 0.69, 0.70, 0.69, and 0.68, respectively. Phase 4 compared these results with existing data. Future prospects that can be considered include:

- Using the Google Maps API instead of the HERE Maps API to download the daylight images;
- Using DHS Data for 2019-20 for more refined results.

Acknowledgements

We would like to thank HERE Maps for providing the dataset for daylight satellite images and for their guidance on how to access these images for training of our model. We are also grateful to the DHS Program and the DMSP-OLS data source for granting us permission to use the DHS Survey data for 2015–2016 and nightlight images, respectively.

References

1. World Data Lab, World Poverty Clock, https://worlddata.io/portfolio/world-poverty-clock/.
2. D. Eugenie, R. Kennedy, J. Urpelainen, Satellite data for the social sciences: measuring rural electrification with night-time lights, *International Journal of Remote Sensing*, **39**(9): 2690–2701, 2018, doi: 10.1080/01431161.2017.1420936.
3. J. Neal, M. Burke, M. Xie, W.M. Davis, D.B. Lobell, S. Ermon, Combining satellite imagery and machine learning to predict poverty, *Science*, **353**(6301): 790–794, 2016, doi: 10.1126/science.aaf789.
4. NOAA, Nightlight Images, DMSP-OLS Dataset, https://ngdc.noaa.gov/eog/dmsp/.
5. The DHS Program, Available datasets, https://dhsprogram.com/data/available-datasets.cfm.
6. K.N. Nischal, R. Radhakrishnan, S. Mehta, S. Chandani, Correlating night-time satellite images with poverty and other census data of India and estimating future trends, [in:] *CODS'15: Proceedings of the 2nd ACM IKDD Conference on Data Sciences*, New York, NY, United States, pp. 75–79, Association for Computing Machinery (ACM), 2015, doi: 10.1145/2732587.2732597.
7. United Nations, A World That Counts: Mobilizing the Data Revolution for Sustainable Development – Report of the Secretary-General's Independent Expert Advisory Group on the Data Revolution for Sustainable Development, 2014, https://digitallibrary.un.org/record/3882725.
8. United Nations, Sustainable Development Goals: 17 Goals to Transform Our World, 2015, http://www.un.org/sustainabledevelopment/sustainable-development-goals/.

9. A. Ajami, M. Kuffer, C. Persello, K. Pfeffer, Identifying a slums' degree of deprivation from VHR images using convolutional neural networks, *Remote Sensing*, **11**(11): 1282, 2019, doi: 10.3390/rs11111282.

10. Daylight Images, Google HERE Maps API, www.here.com.

11. Living Standard Measurement Survey, The World Bank, www.worldbank.org/lsms.

12. E. Sheehan, C. Meng, M. Tan, B. Uzkent, N. Jean, M. Burke, D. Lobell, S. Ermon, Predicting economic development using geolocated Wikipedia articles, [in:] *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2698–2706, Association for Computing Machinery, New York, NY, United States, 2019, doi: 10.1145/3292500.3330784.

13. S. Pandey, T. Agarwal, N.C. Krishnan, Multi-task deep learning for predicting poverty from satellite images, [in:] *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32(1), pp. 7793–7798, New Orleans, Lousiana, USA, 2018, doi: 10.1609/aaai.v32i1.11416.

14. L. Duan, T. Hu, E. Cheng, J. Zhu, C. Gao, Deep convolutional neural networks for spatiotemporal crime prediction, [in:] *Proceedings of the International Conference on Information and Knowledge Engineering (IKE)*, H.R. Arabnia [Ed.], pp. 61–67, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2017.

15. L. Guie, Z. Cai, X. Liu, J. Liu, S. Shiliang, A comparison of machine learning approaches for identifying high-poverty counties: Robust features of DMSP/OLS night-time light imagery, *International Journal of Remote Sensing*, **40**(15): 5716–5736, 2019, doi: 10.1080/01431161.2019.1580820.

16. R. Ruchir, Temporal poverty prediction in developing countries, *Stanford CS229: Machine Learning*, Stanford University, California, 2017, https://cs229.stanford.edu/proj2017/final-reports/5244050.pdf.

17. S.P. Subash, R.R. Kumar, K.S. Aditya, Satellite data and machine learning tools for predicting poverty in rural India, *Agricultural Economics Research Review*, **31**(2): 231–240, 2018, doi: 10.5958/0974-0279.2018.00040.X.

18. B. Raj, A simple guide to the versions of the inception network, *Towards Data Science*, May 29, 2018, https://towardsdatascience.com/a-simple-guide-to-the-versions-of-the-inception-network-7fc52b863202 [accessed: Aug. 20, 2020].

19. M. Xie, N. Jean, M. Burke, D. Lobell, S. Ermon, Transfer learning from deep features for remote sensing and poverty mapping, [in:] *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30(1), pp. 3929–3935, AAAI Press, Palo Alto, California USA, 2016, doi: 10.1609/aaai.v30i1.9906.

20. G.R. Watmough, C.L.J. Marcinko, C. Sullivan, K. Tschirhart, P.K. Mutuo, C.A. Palm, J.C. Svenning, Socioecologically informed use of remote sensing data to predict rural household poverty, [in:] *Proceedings of the National Academy of Sciences*, Vol. 116(4), pp. 1213–1218, 2019, doi: 10.1073/pnas.1812969116.

21. A. Perez, S. Ganguli, S. Ermon, G. Azzari, M. Burke, D. Lobell, Semi-supervised multitask learning on multispectral satellite images using Wasserstein generative adversarial networks (GANs) for predicting poverty, *arXiv*, 2019, arXiv: 1902.11110.

22. S. Piagessi *et al.*, Predicting City Poverty Using Satellite Imagery, [in:] *Proceedings of the IEEE/CFV Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 90–96, Long Beach, California, USA, 2019.

23. J.E. Steele *et al.*, Mapping poverty using mobile phone and satellite data, *Journal of the Royal Society Interface*, **14**(127): 20160690, 2017, doi: 10.1098/rsif.2016.0690.

24. G. Catamuro, A. Muhebwa, J. Taneja, Street smarts: measuring intercity road quality using deep learning on satellite imagery, [in:] *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies, Association for Computing Machinery*, pp. 145–154, New York, NY, United States, 2019, doi: 10.1145/3314344.3332493.

25. P.K. Suraj, A. Gupta, M. Sharma, S.B. Paul, S. Banerjee, On monitoring development indicators using high resolution satellite images, arXiv, 2017, arXiv: 1712.02282.

26. ImageNet, Database, http://www.image-net.org.

27. S. Bhattacharyya, Ridge and Lasso regression: L1 and L2 regularization, *Towards Data Science*, 2018, https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b.

28. RANSAC Regression, Regression Models, 2018, https://scikit-learn.org/stable/modules/classes.html.

29. KNN Nearest Neighbors Regression, 2018, https://scikit-learn.org/stable/supervised_learning.html#supervised-learning.

30. Pearson's Correlation Coefficient, Correlation Coefficient, 2018, https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

31. E. Sheehan, Z. Nabulsi, C. Meng, Utilizing Latent Embeddings of Wikipedia Articles to Predict Poverty, Preprint, Stanford University, 2018, https://cs229.stanford.edu/proj2018/report/134.pdf [accessed: Aug. 20, 2020].

Received July 16, 2022; accepted August 21, 2022;
published online June 24, 2024.

Journal of Optics

ACCEPTED MANUSCRIPT

Pressure-dependent bandgap characteristics in photonic crystals with sensing applications

To cite this article before publication: Varnam Sherawat *et al* 2024 *J. Opt.* in press https://doi.org/10.1088/2040-8986/ad5b73

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is © **2024 IOP Publishing Ltd**.

During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.

As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript will be available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by-nc-nd/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions may be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 14.139.251.98 on 01/07/2024 at 05:42

Pressure-dependent bandgap characteristics in photonic crystals with sensing applications

Varnam Sherawat [a], Renuka Bokolia [b], Ravindra Kumar Sinha[*a,c]

[a] TIFAC-Centre for Relevance and Excellence in Fiber Optics and Optical Communication, Department of Applied Physics, Delhi Technological University, Delhi, India

[b] Department of Applied Physics, Delhi Technological University, Delhi, India

[c]Gautam Buddha University, Greater Noida, Gautam Budh Nagar, UP-201312, India

*Corresponding author: dr_rk_sinha@yahoo.com

Abstract

The present study elucidates a photonic crystal-based pressure sensor exploiting the change in refractive index with pressure and the corresponding structural deformation of the dielectric material. The stress-sensitive refractive indices of the constituent materials of the photonic crystal have been considered to study the effect of applied pressure on the photonic bandgap characteristics of the structure. The designed pressure sensor, proposed using a two-dimensional hexagonal lattice arrangement of air holes in a dielectric slab, operates in the high-pressure range of 1-6 GPa. A comparative study of the photonic bandgap characteristics with the application of high pressure has been reported for three semiconducting materials- GaAs, Ge and Si, used for the dielectric slab in the proposed structure. GaAs is found to exhibit the highest sensitivity to pressure variations and shows more pronounced shifting of the midgap wavelength with pressure in comparison to Ge and Si. The largest photonic bandgap is seen in the Ge-based structure, closely followed by the GaAs and Si-based structures. The proposed structure is suitable for high-pressure sensing applications.

Keywords

Pressure, photonic bandgap (PBG), gap-to-midgap ratio, sensor, photonic crystal (PhC)

1. Introduction

Optical sensors have shown supremacy over conventional sensors owing to the crucial advantages that they offer [1]**.** Optical sensors have exhibited great capacity in sensing physical quantities such as chemical concentration [2], humidity [3], gases [4], bio-analytes [5] etc. Photonic crystals (PhCs) have been used for various sensing applications

Accepted Manuscript

like gas sensing [6], chemical sensing [7], temperature sensing [8,9], biosensing [9–11], etc. due to their high sensitivity over the slightest change in refractive index. Photonic crystals (PhCs) are characterized by one, two or three-dimensional periodic variations in the dielectric constant on the scale of an optical wavelength. The trapping of light within a PhC structure is driven by the photonic bandgap (PBG) effect, which prevents a specific spectral band from propagating through the structure. The tunable PBG makes PhCs more appealing for manipulation of the light propagation characteristics and the radiation fields. Notably, introducing a defect layer disrupts the periodicity, resulting in the creation of strongly localized allowed states within the PBG [12]. The PBG functions as a reflection band, causing the defect mode to manifest as a sharp passband within the PBG [13,14].

PhC sensors are lucrative due to their high sensitivity, low response time, small size, no-label detection, anti-interference ability, high resolution and easy integration with their optical and electronic counterparts [14,15]. PhCs have also been commonly used for mechanical sensing. Force [16–18], displacement [19] and strain sensors [20] based on PhCs have been widely reported. Another form of mechanical sensing where PhCs are used is pressure sensor [21]. Photoelasticity gives a competent tool to design pressure sensors, where the application of stress or pressure leads to change in refractive index and consequently the optical performance of the material [22].

Electrostrictive stress is linked to electric field components through photo-elastic constants, which are proportional to the fourth power of the refractive index. This relationship makes materials with high refractive indices, such as silicon (Si), germanium (Ge), and gallium arsenide (GaAs), particularly significant for optomechanical applications[23,24]. Moreover, using high refractive index materials can also help reduce radiation losses [25]. Additionally, the high refractive index offered by GaAs, Si and Ge provides a considerable refractive index contrast in the air holes in a slab type of PhC structure. Due to this high index contrast, the structure offers a large photonic bandgap which also makes it suitable for wide band applications.

Further, Si, Ge and GaAs are key the intrinsic semiconductors extensively used in electronic devices. Silicon is crucial for most electronic circuits due to its cost-effectiveness, supported by the IC industry, and its superior properties, including high thermal conductivity, an optical damage threshold ten times greater than that of GaAs, a Raman effect 1000 times stronger, and a Kerr effect 100 times larger than that of silica fiber [26,27]. Additionally, silicon is favored for its low-loss waveguides, uniformity, and compact size in photonic devices [28]. GaAs, the second most common semiconductor, is utilized in laser diodes, solar cells, and microwave frequency integrated circuits, boasts of superior

Accepted Manuscript

electrical properties such as higher electron saturation velocity and a direct bandgap. These attributes result in less noise at high frequencies and higher power operation compared to Si [29]. Nevertheless, Si remains dominant due to its abundance, natural oxide (SiO_2) insulator, and ability to form CMOS gates with both P and N transistors, while GaAs is preferred for pressure sensing applications due to its higher sensitivity and better sensing parameters [30]. Advancements in miniaturization and high dielectric systems have made Ge a viable replacement for Si as a channel material [31], offering benefits such as direct CMOS compatibility[32], a high refractive index (~4), and minimal dispersion over a wide temperature range. These properties make Ge excellent for photodetectors in on-chip data distribution [33,34] and valuable for emitters[35], lasers [36], and solar cells. Additionally, existing literature shows that ultrafast optical reflectivity measurements of Si, Ge and GaAs have been carried out using sub-picosecond pulses as pump and broadband 340-780 nm ultrafast pulses as probe [37]. Such intriguing results open avenues for cutting-edge research on photonic devices based on these semiconductors.

Pressure sensors have been reported which work in the megapascals (MPa) or lesser range [38–42]. These sensors are based on the principle of cantilever and higher pressure applied on these leads to permanent deformation. The other class of pressure sensors exploit the variation in refractive index of the PhC material due to applied pressure, leading to shift in the resonant wavelength. Several two-dimensional PhC sensors have been reported in literature which demonstrate shift in resonant wavelength with increase in pressure [30]. These sensors can operate in a comparatively high-pressure range (i.e., GPa).

Table 4 summarizes various PhC structures employed for pressure and force sensing applications, highlighting their design specifications, sensing ranges, sensitivities and simulation platforms. Triangular and square lattice PhCs of dielectric rods in air, as well as 2D PhCs with holes or defect rings, are commonly used, often simulated using finite-difference time-domain (FDTD) or finite element method (FEM) platforms. The PhC-based micro-ring resonator (MRR) structures and hexagonal/triangular lattice hole type PhCs show considerably high sensitivities across different force and pressure ranges. Notably, these PhC based sensor designs need to have a finite structures design with specific boundary conditions in order to deliver the reported results. The plane wave expansion method (PWEM) offers a simpler way to numerically analyse the photonic bandgap with variations in pressure. The PWE simulations, available through means of open source softwares like MPB, assume an infinite PhC structure without the need for any defects/ resonant cavities. This propels the need for a simple and easy to fabricate PhC structure analysed using PWE which

Accepted Manuscript

exploits high index materials – GaAs, Ge, Si; with a comparative study of their PBG properties based on the electrical, semiconducting, optical and photo-elastic properties of the three materials.

The proposed study theoretically analyzes the pressure dependence of the bandgap characteristics and midgap wavelength of a hexagonal lattice photonic crystal of air holes in a semiconducting slab. The study draws a comparison on the effect of pressure variations of the order of GPa on the PhC structure for three semiconductors- GaAs, Si and Ge- used as the dielectric slab. The stress-optic effect, wherein changes in refractive index occur due to application of pressure, has been exploited for the three semiconductors. The changes in the gap-to-midgap ratio and the midgap wavelength with variation in applied pressure have been analyzed for the GaAs, Si and Ge-based PhC structures. GaAs shows the highest-pressure sensitivity owing to maximum change in its refractive index per unit change in pressure, followed by Ge and the least sensitive to pressure variations being Si. The largest PBG is exhibited by Ge followed by GaAs, with Si showing a smaller bandgap than both Ge and GaAs. The reliable and easy to fabricate design positions the proposed structure as a highly promising approach for high pressure sensing applications.

2. Theoretical Design

The proposed structure consists of air holes in a dielectric slab, designed in a hexagonal lattice arrangement. The lattice constant (a) for proposed arrangement has been taken as 485 nm with the radius of air holes (r) taken as 0.45a. The optimization of these parameters is conducted to obtain a large bandgap approximately within the wavelength range of 1300 nm to 1400 nm for the three semiconductors- GaAs, Si and Ge- used for the slab. The precise values of the geometrical parameters are optimized so that the GaAs based PhC yields maximum PBG at the maximum value of applied pressure (6 GPa) such that midgap wavelength at this pressure is ~ 1350 nm. The optimized parameters have been listed in Table 1.The proposed PhC structure has been depicted in Fig. 1. A hexagonal lattice arrangement is used since it can give a bandgap for both TM and TE polarizations. This is because the hexagonal arrangement provides a structure with high-dielectric regions that are both practically isolated and linked by narrow veins [8]. This leads to a connected structure which has channels along which the field can propagate [43].

Accepted Manuscript

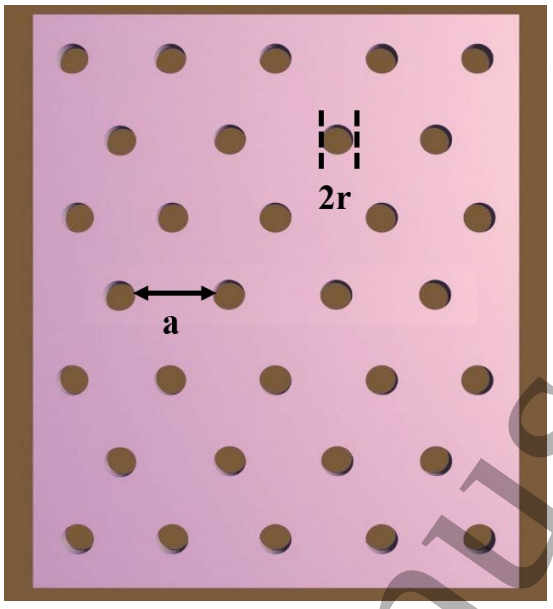


Fig. 1. Schematic of the proposed PhC with hexagonal lattice

Apart from the fact that the proposed design is an infinite PhC structure without any defects, it should also be pointed out that Si, Ge and GaAs have different refractive indices and thus offer a different value of index contrast in the proposed air hole in slab type of geometry. Hence, all the three materials yield maximum PBG at different wavelengths. This leads to the assumption of an approximate optimization wavelength (~1350 nm) common to all the three materials for the purpose of an objective comparison.

S. No.	**Parameters**	**Designation**
1.	Lattice structure	Hexagonal lattice
2.	Configuration	Air holes in dielectric slab
3.	Lattice constant (a)	485 nm
4.	Radius of the air holes (r)	0.45*a
5.	RI of slab (without the applied pressure)	$n_{GaAs} \sim 3.43$; $n_{Ge} \sim 4$; $n_{Si} \sim 3.42$
6.	Operating wavelength range (for all the three-semiconductor based PhCs)	1290-1400 nm

Accepted Manuscript

7.	Optimisation wavelength	~1350 nm
8.	Computational method	Plane Wave Expansion (PWE)
9.	Simulation platform	MIT- Photonic Bands (MPB)
10.	Defects and Dimensions	No defects; infinite PhC structure (for PWE based PBG calculations in MPB)

Table 1. Design specifications of the proposed air holes in slab type PhC sensor

2D PhCs can provide a full PBG (not in a strict sense), not possible in case of 1D PhCs which are limited to one-dimension in terms of the dielectric periodicity, and they are simpler to fabricate than 3D PhCs [44]. The air holes in slab kind of structure has very low losses [30]. Hence, the proposed 2D PhC is appealing for the current research work. The proposed design is reliable and simple to be integrated into various detection applications and optical devices such as filters, splitters, cavities, multiplexers/demultiplexers, sensors, etc [45]. This hole type PhC structure has better pressure detection ability and easier to fabricate compared to the rod type structure [46] and is useful for high pressure measurement applications. The 2D PhC-based structure can be realized by etching holes in the higher dielectric material periodically in two dimensions. The etching methods have been used for semiconductors, where the marking of unwanted areas is done on the surface of a semiconductor, using a lithographic technique such as electron beam lithography. The marked areas are then etched to create holes [47].

3. Pressure dependance of refractive index

The stress-strain relation and the photo-elastic effect dictate the relation between refractive index and stress, which takes a tensor form as follows [48]:

$$\boldsymbol{\Delta}\left(\frac{\mathbf{1}}{\mathbf{n^2}}\right) = \sum \mathbf{p_{ij}}\,\boldsymbol{\sigma_j} \qquad (\mathbf{1})$$

Here, $\mathbf{n}$ is the refractive index of the slab material, $\mathbf{p_{ij}}$ are the strain-optic coefficients and $\boldsymbol{\sigma_j}$ is the component of stress. Following the assumption that the pressure is applied in one direction only with the medium being isotropic, equation $(\mathbf{1})$ takes a simplified form [49]:

$$\mathbf{n} = \mathbf{n_0} - (\mathbf{c_1} + \mathbf{2c_2})\boldsymbol{\sigma} \qquad (\mathbf{2})$$

Accepted Manuscript

Where, $\mathbf{n}$ is the refractive index of the sensor plate matrix and $\mathbf{n_0}$ is the refractive index of the slab without application of pressure. $\mathbf{c_1}$ and $\mathbf{c_2}$ are the strain-optic constants defined as [49]:

$$\mathbf{c_1} = \frac{\mathbf{n_0^3}\,(\mathbf{p_{11}} - \mathbf{2Vp_{12}})}{\mathbf{2E}} \quad (3)$$

$$\mathbf{c_2} = \frac{\mathbf{n_0^3}\,[\mathbf{p_{12}} - \mathbf{V}(\mathbf{p_{11}} + \mathbf{p_{12}})]}{\mathbf{2E}} \quad (4)$$

Here, $\mathbf{p_{11}}$ and $\mathbf{p_{12}}$ are the strain-optic coefficients, $\mathbf{V}$ is the Poisson's ratio and $\mathbf{E}$ is the Young's modulus. The strain-optic constants for GaAs and Si are given in Table 2 [50]. Thus, GaAs and Si exhibit a stress-optic effect which is linear, as shown in Fig. 2.

Material	$\mathbf{n_0}$	$\mathbf{c_1}$ ($\mathbf{10^{-12}/Pa}$)	$\mathbf{c_2}$ ($\mathbf{10^{-12}/Pa}$)
Si	3.42	-11.35	3.65
GaAs	3.43	-18.39	-10.63

Table 2. Strain-optic constants

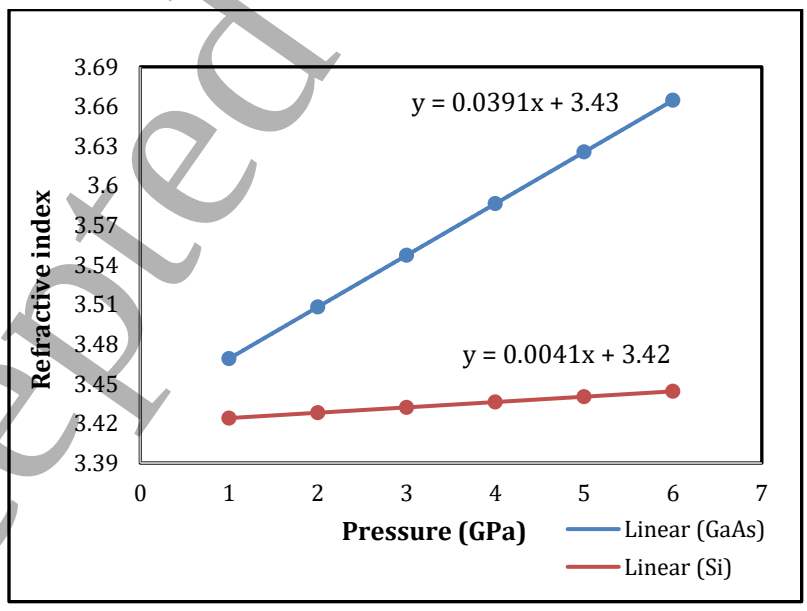


Fig. 2. Refractive index profile of GaAs and Si with variation in applied pressure

Accepted Manuscript

Another material exhibiting the stress-optic effect is germanium (Ge). The variation of refractive index of Ge n_{Ge} on application of hydrostatic pressure P to its surface [51] is given by:

$$n_{Ge} = \sqrt{15.94 - 0.36P + 0.014\,P^2} \quad (5)$$

This relation is used to obtain the variation of refractive index of Ge with pressure varying in the range 0 – 10 GPa, as shown in Fig. 3.

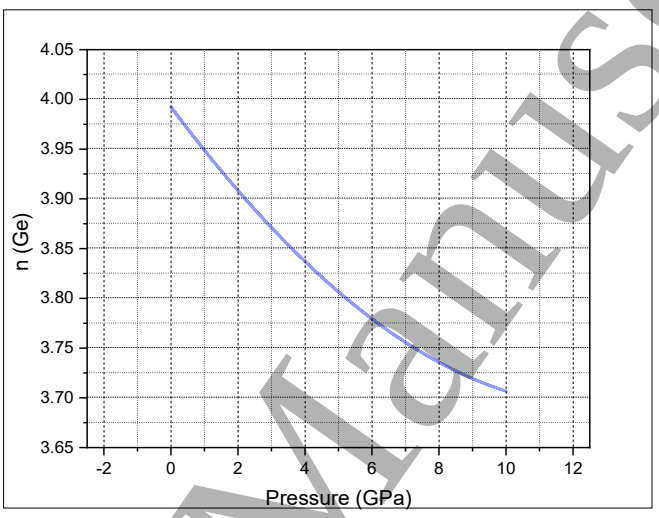


Fig. 3. Refractive index of Ge as a function of pressure

4. Results and Discussion

The MIT Photonic Bands (MPB) software has been employed to compute the frequency eigenstates of the Maxwell's equation using the plane wave expansion (PWE) method. The eigen-frequencies for each wavevector (or k-point) gives the band structure of the proposed PhC design. The photonic band structure calculations have been performed for the three different semiconducting materials- GaAs, Ge and Si, used as the dielectric slab keeping all the structural parameters unchanged. The output frequency values are normalized in units of c/a, where c and a represent the speed of light and lattice constant of the structure respectively. The corresponding wavelengths are determined by employing the optimized value of the lattice constant in the normalized output. The design and geometry have been optimized to obtain wide bandgaps, since large band structures find utility in sensing applications.

Accepted Manuscript

4.1 GaAs

The bandgap calculations performed for the proposed PhC design with the dielectric slab composed of GaAs show the existence of a wide photonic bandgap (PBG) for the TE mode (~ 48.87 %) and a much smaller PBG for the TM mode (~ 9.71 %). The calculations have been performed for various values of pressure in the range 1 GPa - 6 GPa. For this pressure range, the TE bandgap, gap-to-midgap ratio and the band-edge wavelengths have been computed for different values of applied pressure. The GaAs based structure shows are a large bandgap for the TE mode ~ 48.87 % at 1GPa pressure between bands 1 and 2, as can be seen from Fig. 4.

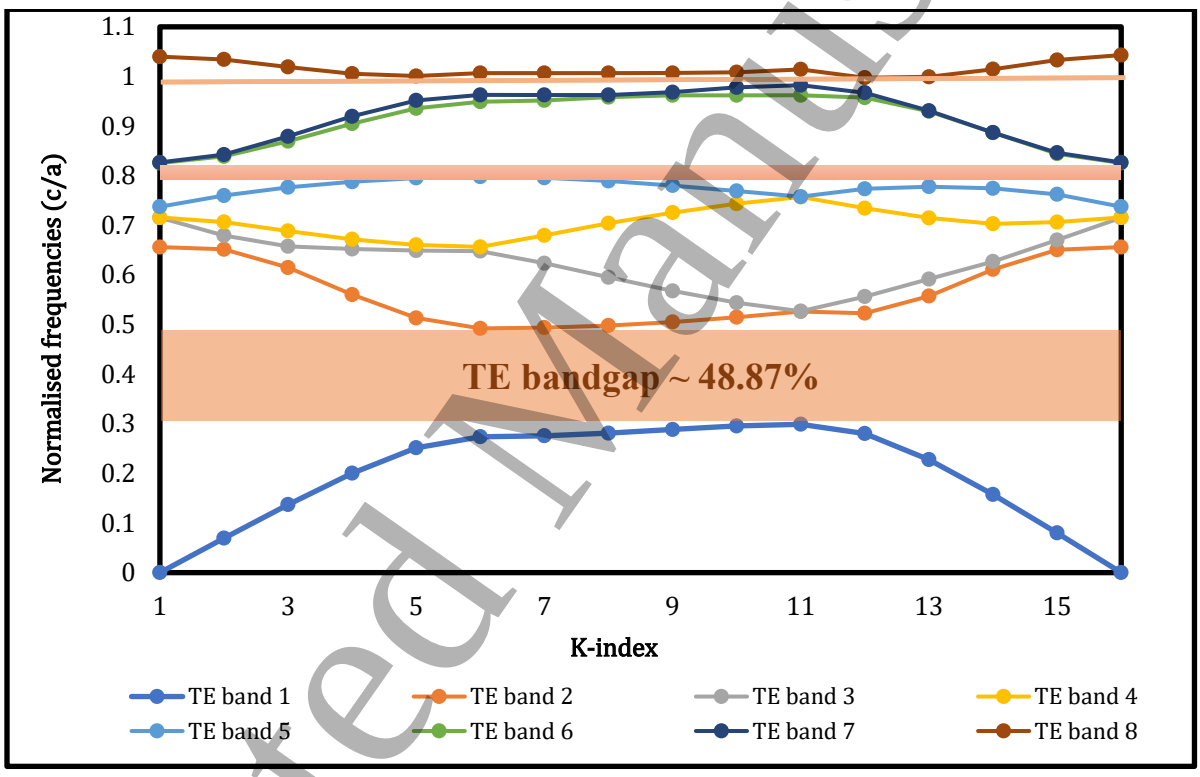


Fig. 4. Photonic band structure of the GaAs-based PhC for the TE mode at 1GPa pressure

The central wavelength of the photonic bandgap, i.e., the midgap wavelength of the structure, has been evaluated at different values of applied pressure P in the range 1 GPa -6 GPa, using the corresponding band-edges, as per the following equation:

$$\lambda(\mathrm{P})_{\mathrm{midgap}} = \frac{\lambda(\mathrm{P})_{\mathrm{upper\ band\ edge}} + \lambda(\mathrm{P})_{\mathrm{lower\ band\ edge}}}{2} \quad (6)$$

The midgap wavelength increases linearly with increase in applied pressure, as depicted in Fig. 5. Hence, the midgap wavelength is sensitive to changes in applied pressure and can be tuned with pressure. The pressure

Accepted Manuscript

sensitivity, S_p , is calculated as the ratio of the change in midgap wavelength ($\Delta\lambda$) to the change in applied pressure (ΔP) , measured in units of nm/GPa.

$$S_p = \frac{\Delta\lambda}{\Delta P} = \frac{\Delta\lambda}{\Delta n}\frac{\Delta n}{\Delta P} \tag{7}$$

The GaAs-based PhC structure shows a pressure sensitivity of 9.294 nm/GPa and can be used for high pressure sensing applications. Here, $\Delta\lambda/\Delta P > 0$ i.e., the central wavelength increases for every 1 GPa increase in pressure.

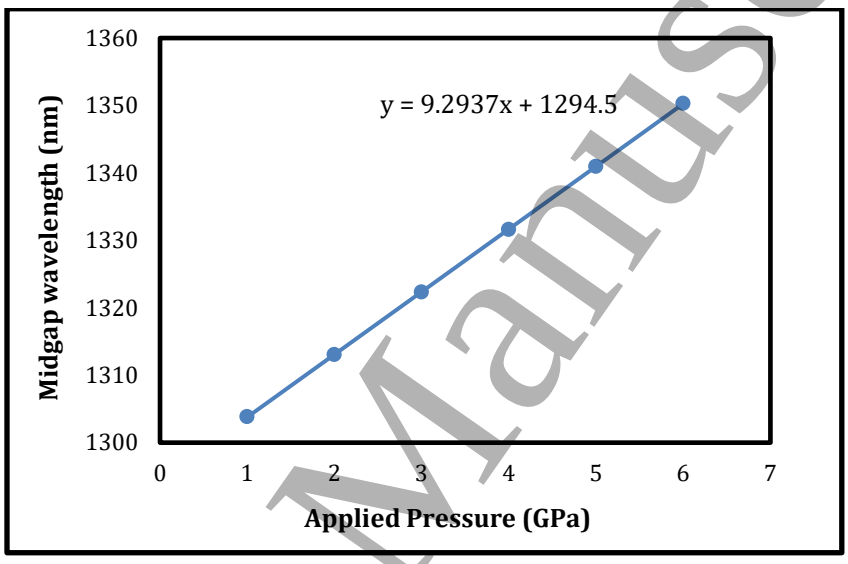


Fig. 5. Midgap wavelength of the GaAs-based PhC as a function of applied pressure

The gap-to-midgap ratio, determined for different values of applied pressure in the range 1 GPa – 6 GPa, has been shown in Fig. 6. It is seen to increase linearly with increase in pressure. This is expected, owing to the fact that the refractive index of the GaAs slab increases with increasing pressure (as seen in Fig. 2 as per equation (2)). This increase in refractive index leads to an increase in the dielectric contrast between the slab and the air holes of the PhC structure, causing the PBG to increase with applied pressure. This consequently is responsible for the increase in gap-to-midgap ratio with applied pressure.

Accepted Manuscript

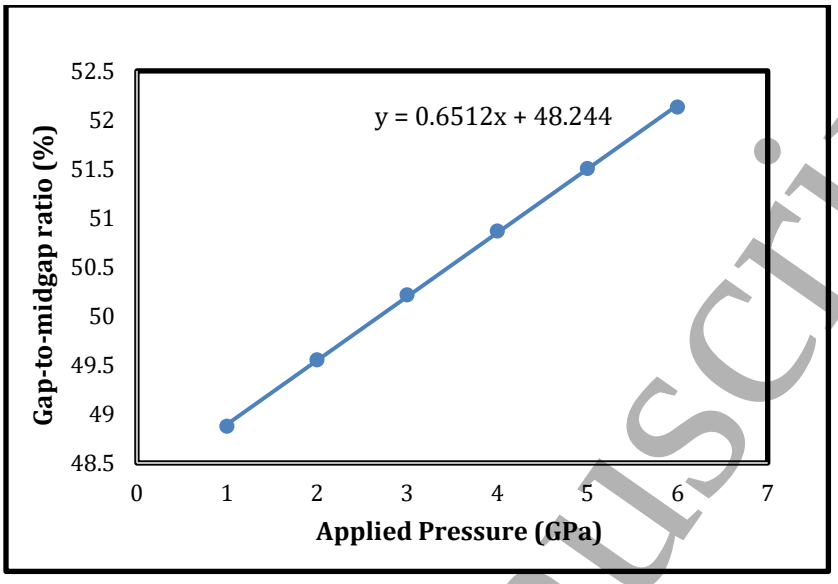


Fig. 6. Change in the gap-to-midgap ratio with variation of applied pressure for the GaAs-based PhC

4.2 Ge

The bandgap calculations for the Ge-based PhC reveal the existence of a large TE bandgap (~ 56.31 %) between bands 1 and 2 along with a small TM bandgap (~ 11.76 %) between bands 2 and 3 at 1 GPa pressure. The photonic band structure along with the bandgap for the TE mode has been shown in Fig. 7.

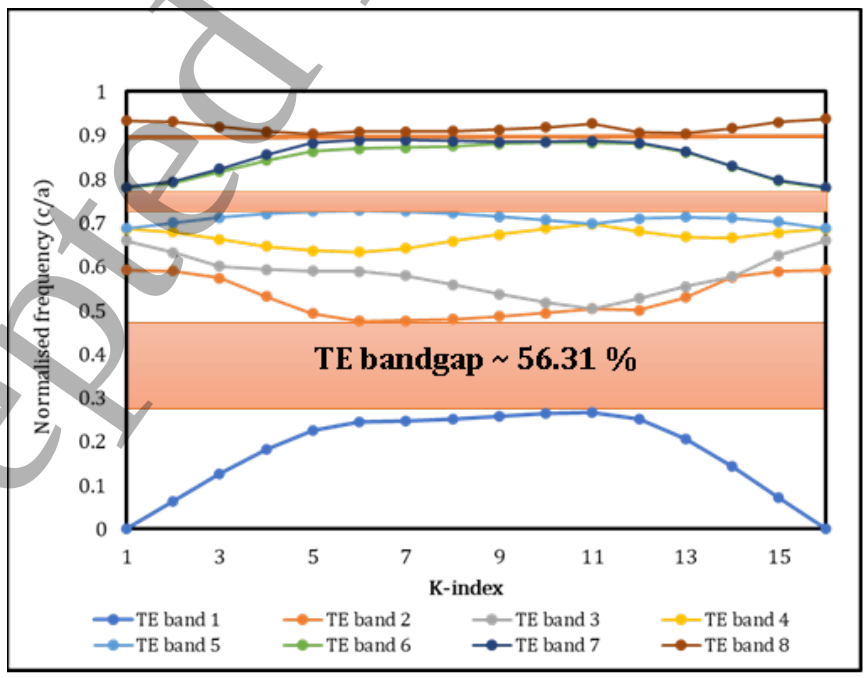


Fig. 7. Photonic band structure of the Ge-based PhC for the TE mode at 1GPa

Accepted Manuscript

The midgap wavelength of the Ge-based PhC is determined for different pressure values in the range 1 GPa to 6 GPa as per equation (**6**) by using the corresponding values of the band-edge wavelengths as at that particular pressure. The midgap wavelength exhibits a linear decrease with increase in pressure as shown in Fig. 8. The pressure sensitivity of the structure, calculated as per equation (**7**), is observed to be -8.295 nm/GPa. Here, the negative sign is indicative of the decrease in the central wavelength for every 1 GPa increase in pressure (i.e., $\boldsymbol{\Delta\lambda/\Delta P < 0}$).

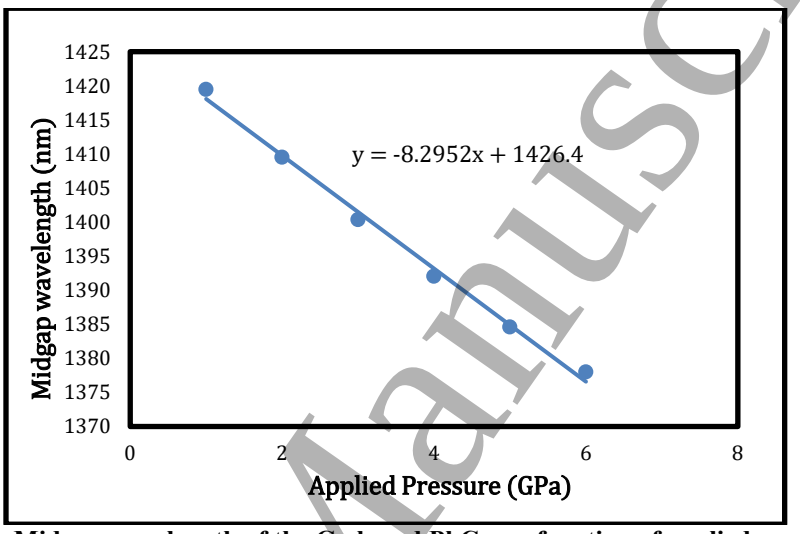


Fig. 8. Midgap wavelength of the Ge-based PhC as a function of applied pressure

The variation of the gap-to-midgap ratio with pressure, applied in the range 1 GPa – 6 GPa, has been shown in Fig. 9 for the Ge-based PhC. The gap-to-midgap ratio decreases linearly with increase in applied pressure. This is because the refractive index of the Ge slab decreases with increasing pressure (as seen in Fig. 3 as per equation (**5**)). This decrease in refractive index causes a decrease in the dielectric contrast between the slab and the air holes leading to a decrease in the PBG and the gap-to-midgap ratio of the Ge-based PhC.

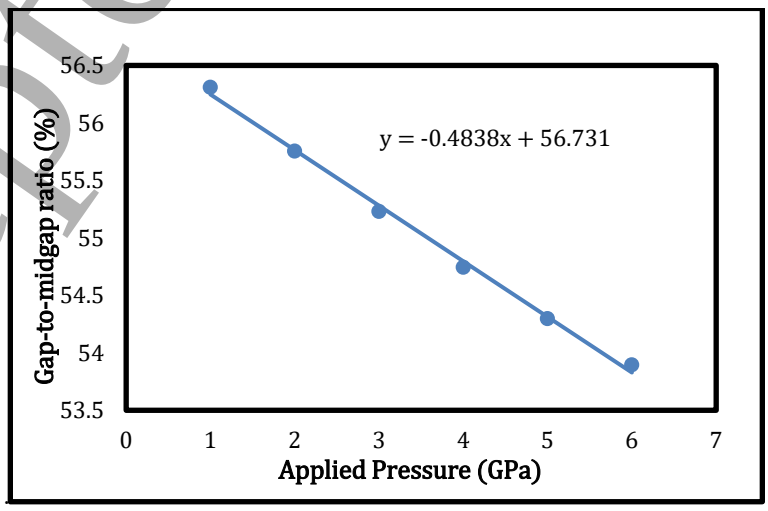


Fig. 9. Change in the gap-to-midgap ratio with applied pressure for the Ge-based PhC

Accepted Manuscript

4.3 Si

The PBG calculations performed for the Si-based PhC shows the existence of a large TE bandgap (~ 48.07 %) between bands 1 and 2 and a much narrower gap for the TM mode (~ 9.47 %) between bands 2 and 3. The TE bandgap for the Si-based PhC is depicted in the TE band structure diagram in Fig. 10.

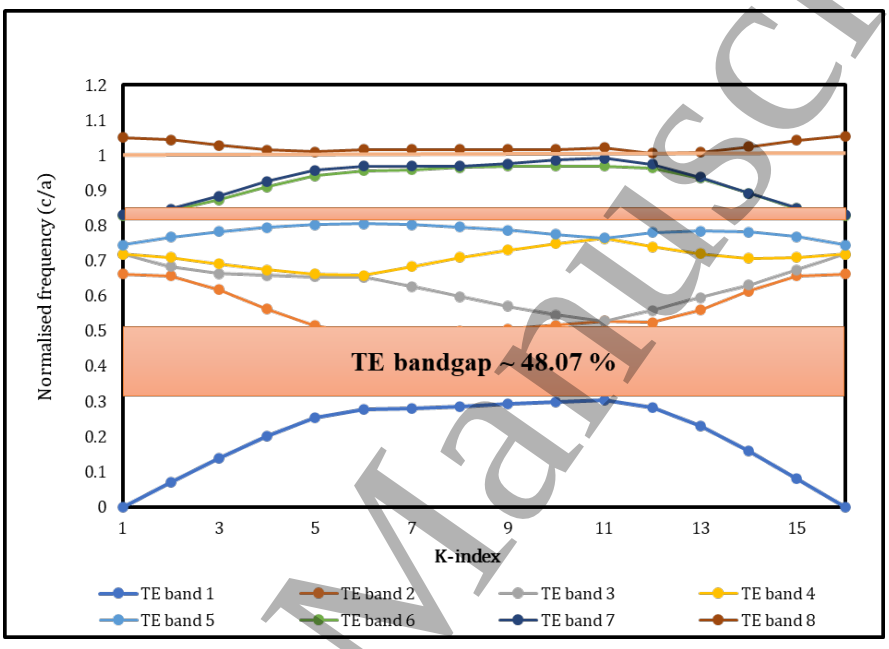


Fig. 10. Photonic band structure of the Si-based PhC for the TE mode at 1GPa pressure

For different values of pressure in the range 1 GPa – 6 GPa, the midgap wavelength of the Si-based structure is calculated as per equation (6). The midgap wavelength is observed to increase linearly with increase in applied pressure as seen in Fig. 11. The pressure sensitivity of the structure, as per equation (7), is calculated to be 0.9502 nm/GPa. The positive value of the slope in Fig. 8, i.e., $\Delta\lambda/\Delta P > 0$ is indicative of the increase in the central wavelength of the TE PBG for every 1 GPa increase in applied pressure.

Accepted Manuscript

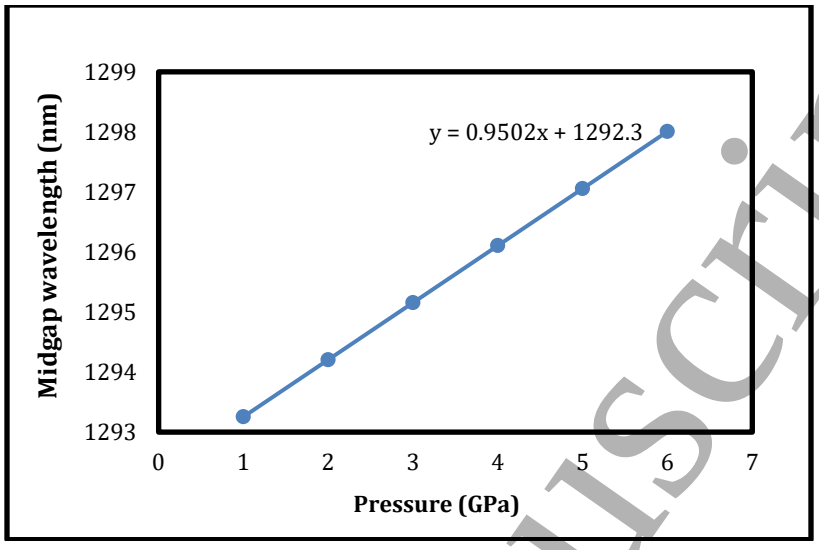


Fig. 11. Midgap wavelength of the Ge-based PhC as a function of applied pressure

The gap-to-midgap ratio of the Si-based PhC determined for various values of applied pressure from 1 GPa – 6 GPa shows a linear increase with pressure as seen from Fig. 12. This is owing to the increase in refractive index of the Si slab with increase in applied pressure. This causes an increase in the refractive index contrast between the Si slab and the air holes, leading to an increase in the bandgap, which in turn is responsible for the increase in the gap-to-midgap ratio of the structure.

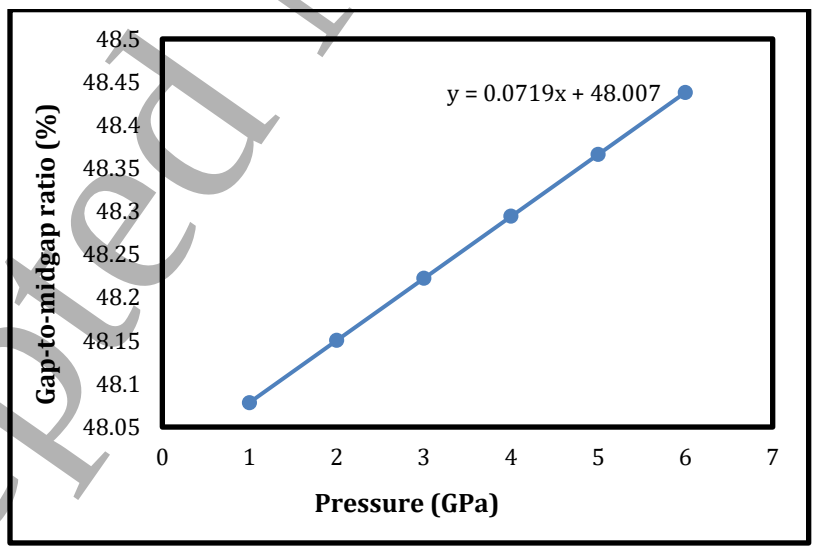


Fig. 12. Gap-to-midgap ratio as a function of applied pressure for the Ge-based PhC

Accepted Manuscript

4.4 Comparative analysis of the GaAs, Ge and Si-based structures

The proposed PhC design has been realized using three different semiconducting materials- Si, Ge and GaAs for the dielectric slab. Si and Ge are monocrystalline semiconductors i.e., they have a repeating crystal structure. On the other hand, GaAs is a compound semiconductor i.e., composed of one or more semiconducting materials with different atomic structures.

Semiconducting slab material	**Young's Modulus (E)**	**Δ n for Δ P = 1GPa**	**Pressure sensitivity (S_p)**	**Normalised sensitivity (S_n)**	**Refractive index at 1 GPa**	**PBG at 1 GPa**
GaAs	85.5 GPa	0.0391	9.294 nm/GPa	0.006884222 /GPa	3.4691	48.87%
Ge	100 GPa	0.0355	-8.295 nm/GPa	0.006144444 /GPa	3.9489	56.31 %
Si	165 GPa	0.0041	0.9502 nm/GPa	0.000703852 /GPa	3.4241	48.07 %

Table 3. Parameters for comparisons of the three PhC structures

The refractive index of GaAs and Si increases linearly with increase in pressure whereas the refractive index of Ge exhibits a non-linear decrease with increase in pressure as seen in Fig. 2 and Fig. 3. The pressure applied induces a certain strain and stress in the material. The major reason for a significant change in the refractive index of the material is the strain-induced change of the energy bandgap or the absorption coefficients [52,53].

The calculated pressure sensitivity is lowest for Si and highest for GaAs as seen from Table 3. This is a consequence of the fact that the change in refractive index for 1 GPa change in pressure is highest for GaAs and lowest for Si among the three semiconducting materials as shown in Table 3. The Young's modulus of Si is the highest and the of GaAs is the lowest among the three materials. Hence, Si offers the most rigid structure with the least deformations as compared to the other two structures. GaAs, on the other hand, exhibits more strain for the same value of applied pressure. This strain-induced change of energy bandgap causes change in the refractive

Accepted Manuscript

index which is most significant in GaAs. Thus, GaAs shows the highest variation of refractive index with pressure and consequently, is the most sensitive to changes in applied pressure.

For a more objective comparison of sensitivity of the proposed PhC sensors, normalised sensitivity has been evaluated. The normalised sensitivity is defined as ratio of the pressure sensitivity to the operating wavelength of the defect mode. Since the proposed structure is an infinite PhC without any defects, the operating wavelength is taken to be the average wavelength (~1350 nm approx.) of the wavelength range (~ 1300 -1400 nm) around which the PhC's geometrical parameters have been optimized for the three semiconductors. Additionally, due to the different RIs of GaAs, Si and Ge, they offer maximum PBG at different wavelengths. This leads to the assumption of an approximate optimization wavelength common to all the three materials for the purpose of comparison. Hence, the normalised sensitivity (S_n) has been calculated and tabulated in Table 3 using the following equation:

$$S_n = \frac{\text{Pressure sensitivity}}{\text{Operating wavelength}} = \frac{S_P}{\lambda_{optimisation}} \tag{8}$$

where, $\lambda_{optimisation}$ is approximated to 1350 nm for normalised sensitivity calculations. As is expected on the basis of the pressure sensing performance, the normalised sensitivity is highest for the GaAs based PhC sensor, followed by Ge, with the Si based PhC exhibiting the lowest sensitivity to pressure variations as interpreted from Table 3.

GaAs is a direct bandgap semiconductor having a zincblende structure whereas Si and Ge are indirect bandgap semiconductors with a diamond structure. This causes GaAs to have a greater electronic bandgap than Si and Ge. Being indirect bandgap semiconductors, Si and Ge need phonon emission or absorption, hence the efficiency of a device based on these will be effectively lesser than that of GaAs. Due to greater resistance of GaAs as compared to Si and Ge, there are a large number of localized states available in GaAs. Therefore, on application of pressure the atoms have larger number of possible states that they can occupy in GaAs. As pressure is applied, the atoms acquire a configuration in which they are closer to each other and hence the refractive index increases considerably [54]. The greater increase in refractive index leads to larger shifts in the midgap wavelength for the GaAs-based PhC. Si, on the other hand, has fewer localized states available and therefore, shows a lesser increase in refractive index. This is responsible for the greater sensitivity of the GaAs based PhC pressure sensor as

Accepted Manuscript

compared to Ge and Si. The largest photonic bandgap, however, is seen in the Ge-based PhC structure, followed by the GaAs and then the Si-based PhC structures. This is a consequence of the fact that Ge shows the largest change in refractive index for every 1 GPa change in pressure applied to its surface as compared GaAs and Si (as seen from Table 3).

A comparative study has been conducted to draw a contrast between the pressure sensing performances of the proposed PhC sensor and similar PhC based pressure sensing structures previously reported in literature. The comparative analysis has been summarized in Table 4. It can be observed that the proposed work is the only one that employs PWE and doesn't contain any defects/resonant cavity. Yet it offers a considerably high sensitivity over a good enough sensing range. This hole type PhC structure has better pressure detection ability compared to the rod type structure [46]. It is is easy to fabricate, reliable and simple to be integrated into various high pressure detection applications.

Reference	**Sensing Range**	**Sensitivity**	**Proposed structure**	**Simulation platform**	**Applic-ation**
[55]	0-4 GPa	1.5 nm/GPa	Triangular lattice of rods in air	FDTD	PS
[56]	0-7 GPa	2 nm/GPa	Square lattice of rods in air	FDTD	PS
[57]	0-5 GPa	2 nm/GPa	2D PhC with square array of holes in slab with 2 WGs and a hexagonal defect ring	FEM	PS
[58]	0-5 GPa	2.19 nm/GPa	Triangular lattice of holes with 2 WGs coupled through a NC	FDTD	PS
[59]	0–1 μN	9.33 nN per 0.1 nm	PhC based MRR structure	FDTD and FEM	FS
[60]	58 nN (HS) and 44 nN (VS)	7.5 nm/μN (HD) and 10 nm/μN (VD)	2D hexagonal lattice PhC	FDTD and FEM	SS

Accepted Manuscript

[61]	Not mentioned	1.891 nm/μN	2D PhC with a NC	FDTD and FEM	FS
[62]	0-2 GPa	13.9 nm/GPa	PhC having 2 WGs and 1 NC	FEM	PS
[46]	0-0.5 GPa (hole type) 0-10 GPa (rod type)	25.26 nm/GPa (hole type) 2.77 nm/GPa (rod type)	2D PhC MRR with square lattice (holes in slab and rods in air)	FDTD and FEM	PS
[63]	0.1-10 GPa	8 nm/GPa	PhC with square array of rods in air with WG	FDTD	PS
Proposed work	1– 6 GPa	9.294 nm/GPa (GaAs PhC) 8.295 nm/GPa (Ge PhC) 0.9502 nm/GPa (Si PhC)	PhC with hexagonal lattice of air holes in SC slab without any defects/cavity	PWE	PS

Table 4. Comparison with previously reported PhC based pressure sensors

Abbreviations: PS (Pressure Sensor); FS (Force Sensor); SS (Stress Sensor); MRR (micro ring resonator); FDTD (Finite-Difference Time-Domain); FEM (finite element method); HS (horizontally applied stress); VS (vertically applied stress); HD (horizontal direction); VD (vertical direction); NC (nanocavity); WG (waveguide); SC (semiconducting); PWE (plane wave expansion);

5. Conclusion

This paper presents an optical platform for pressure sensing based on a hexagonal lattice photonic crystal of air holes in a dielectric slab. The pressure detection characteristics are theoretically analyzed based on the Plane Wave Expansion method. The study draws comparison of the effect of pressure variations on the photonic bandgap characteristics of the proposed PhC design for three semiconductors- GaAs, Ge and Si- used for the dielectric slab. GaAs and Si show an increase in refractive index and hence, increase in the photonic bandgap of their corresponding PhC structures with increase in applied pressure. Ge, on the other hand, shows decrease in its refractive index and

Accepted Manuscript

hence decrease in bandgap with increasing values of pressure. The change in the midgap wavelength and the gap-to-midgap ratio with applied pressure for the three PhC structures reveal that GaAs exhibits the highest-pressure sensitivity of 9.294 nm/GPa, followed by Ge with 8.295 nm/GPa and the least sensitive being Si with a sensitivity of only 0.9502 nm/GPa. The largest PBG is shown by the Ge-based PhC followed by the GaAs and Si based structures. Since the proposed PhC sensor has been analyzed for the range 1- 6 GPa, it is suitable for high pressure sensing applications such as downhole application [64], quantification of pressure levels during the testing of energetic materials [65].

6. Acknowledgements

The authors gratefully acknowledge the initiatives and support provided by the establishment of the TIFAC-Centre of Relevance and Excellence in Fiber Optics and Optical Communication at Delhi Technological University (Formerly Delhi College of Engineering) Delhi, through the "Mission Reach" program of Technology Vision-2020 of the Government of India. One of the authors, Varnam Sherawat (VS), acknowledges the Ph.D. fellowship provided by Delhi Technological University, Delhi, India. She is also grateful to Dr. Yogita Kalra for her valuable suggestions on this research work.

7. References

[1] Regtien P and Dertien E 2018 Optical sensors *Sensors for Mechatronics* (Elsevier) pp 183–243

[2] Memon S F, Wang R, Strunz B, Chowdhry B S, Pembroke J T and Lewis E 2022 A Review of Optical Fibre Ethanol Sensors: Current State and Future Prospects *Sensors* **22** 950

[3] Shao M, Liu B, Wang Y, Liu Y and Qiao X 2022 Optical fiber humidity sensor based on a hollow core fiber filled with BPQDs-PVA *Appl. Opt.* **61** 10439

[4] Bojęś P, Jaworski P, Krzempek K, Malecha Z, Yu F, Wu D, Kozioł P, Dudzik G, Liao M and Abramski K 2022 Experimental and numerical analysis of gas flow in nodeless antiresonant hollow-core fibers for optimization of laser gas spectroscopy sensors *Opt. Laser Technol.* **152** 108157

[5] Ahlawat L, Kishor K and Sinha R K 2024 Photonic spin Hall effect-based ultra-sensitive refractive index sensor for haemoglobin sensing applications *Opt. Laser Technol.* **170** 110183

[6] Goyal A K, Dutta H S and Pal S 2017 Recent advances and progress in photonic crystal-based gas sensors *J.*

Accepted Manuscript

Phys. D. Appl. Phys. **50** 203001

[7] Cai Z, Smith N L, Zhang J-T and Asher S A 2015 Two-Dimensional Photonic Crystal Chemical and Biomolecular Sensors *Anal. Chem.* **87** 5013–25

[8] Sherawat V, Bokolia R and Sinha R K 2022 Impact of thermal and refractive index tuning on the bandgap and band-edges of a silicon photonic crystal waveguide with sensing applications *Opt. Commun.* **518** 128348

[9] Sherawat V, Bokolia R and Sinha R K 2023 Studies on temperature-dependent bandgap and gap-to-midgap ratio in diamond lattice photonic crystal with biosensing applications *Opt. Mater. (Amst).* **145** 114470

[10] Grepstad J O, Kaspar P, Solgaard O, Johansen I-R and Sudbø A S 2012 Photonic-crystal membranes for optical detection of single nano-particles, designed for biosensor application *Opt. Express* **20** 7954

[11] Feng S, Jiang J-H, Rashid A Al and John S 2016 Biosensor architecture for enhanced disease diagnostics: lab-in-a-photonic-crystal *Opt. Express* **24** 12166

[12] Aly A H, Sayed F A and Elsayed H A 2020 Defect mode tunability based on the electro-optical characteristics of the one-dimensional graphene photonic crystals *Appl. Opt.* **59** 4796

[13] Notomi M, Shinya A, Yamada K, Takahashi J-I, Takahashi C and Yokohama I 2002 Structural tuning of guiding modes of line-defect waveguides of silicon-on-insulator photonic crystal slabs *IEEE J. Quantum Electron.* **38** 736–42

[14] Wu C, Liu X, Feng S, Chen X, Li C and Wang Y 2018 High-sensitivity silicon-based photonic crystal refractive index biosensor based on defect-mode coupling *Opt. Commun.* **427** 409–17

[15] Passaro V, Tullio C, Troia B, Notte M, Giannoccaro G and Leonardis F 2012 Recent Advances in Integrated Photonic Sensors *Sensors* **12** 15558–98

[16] Stomeo T, Grande M, Qualtieri A, Passaseo A, Salhi A, De Vittorio M, Biallo D, D'orazio A, De Sario M, Marrocco V, Petruzzelli V and Prudenzano F 2007 Fabrication of force sensors based on two-dimensional photonic crystal technology *Microelectron. Eng.* **84** 1450–3

[17] Li B and Lee C 2010 Computational study of NEMS diaphragm sensor using triple nano-ring resonator *Procedia Eng.* **5** 1418–21

Accepted Manuscript

[18] Li B, Hsiao F-L and Lee C 2011 Configuration analysis of sensing element for photonic crystal based NEMS cantilever using dual nano-ring resonator *Sensors Actuators A Phys.* **169** 352–61

[19] Yang D 2011 Microdisplacement sensor based on high-Q nanocavity in slot photonic crystal *Opt. Eng.* **50** 054402

[20] Tung B T, Dao D V, Ikeda T, Kanamori Y, Hane K and Sugiyama S 2011 Investigation of strain sensing effect in modified single-defect photonic crystal nanocavity *Opt. Express* **19** 8821

[21] Sherawat V, Bokolia R and Sinha R K 2023 Tunable pressure sensor based on two-dimensional silicon photonic crystal *Photonic Fiber and Crystal Devices: Advances in Materials and Innovations in Device Applications XVII* ed S Yin and R Guo (SPIE) p 29

[22] Jeiping Xu R S 1992 Acousto-optic devices : principles, design, and applications *Wiley series in pure and applied optics* ed J W (Standford U Goodman (New York : Wiley)

[23] Rakich P T, Davids P and Wang Z 2010 Tailoring optical forces in waveguides through radiation pressure and electrostrictive forces *Opt. Express* **18** 14439

[24] Baker C, Hease W, Nguyen D-T, Andronico A, Ducci S, Leo G and Favero I 2014 Photoelastic coupling in gallium arsenide optomechanical disk resonators *Opt. Express* **22** 14072

[25] Adawi A M, Chalcraft A R, Whittaker D M and Lidzey D G 2007 Refractive index dependence of L3 photonic crystal nano-cavities *Opt. Express* **15** 14299

[26] Wu J, Peng J, Liu B, Pan T, Zhou H, Mao J, Yang Y, Qiu C and Su Y 2016 Passive silicon photonic devices for microwave photonic signal processing *Opt. Commun.* **373** 44–52

[27] Jalali B and Fathpour S 2006 Silicon Photonics *J. Light. Technol.* **24** 4600–15

[28] Sherawat V, Bokolia R and Sinha R K 2022 Impact of thermal and refractive index tuning on the bandgap and band-edges of a silicon photonic crystal waveguide with sensing applications *Opt. Commun.* **518** 128348

[29] Mostefai A 2022 Comparison Between Silicon (Si) and Gallium Arsenide (GaAs) Using MATLAB *J. Nano-Electron. Phys.* **14** 04028-1-04028–4

[30] Dinodiya S and Bhargava A 2022 A Comparative Analysis of Pressure Sensing Parameters for Two

Accepted Manuscript

Dimensional Photonic Crystal Sensors Based on Si and GaAs *Silicon* **14** 4611–8

[31] Husanu M A, Popescu D G, Ganea C P, Anghel I and Florica C 2015 Correlation of optical reflectivity with numerical calculations for a two-dimensional photonic crystal designed in Ge *Eur. Phys. J. D* **69** 273

[32] Chi On Chui K C S Advanced Germanium MOS Devices and Technology *2005 IEEE Conference on Electron Devices and Solid-State Circuits* (IEEE) pp 101–6

[33] Assefa S, Xia F and Vlasov Y A 2010 Reinventing germanium avalanche photodetector for nanophotonic on-chip optical interconnects *Nature* **464** 80–4

[34] Michel J, Liu J and Kimerling L C 2010 High-performance Ge-on-Si photodetectors *Nat. Photonics* **4** 527–34

[35] Boucaud P, El Kurdi M, Ghrib A, Prost M, de Kersauson M, Sauvage S, Aniel F, Checoury X, Beaudoin G, Largeau L, Sagnes I, Ndong G, Chaigneau M and Ossikovski R 2013 Recent advances in germanium emission [Invited] *Photonics Res.* **1** 102

[36] Camacho-Aguilera R E, Cai Y, Patel N, Bessette J T, Romagnoli M, Kimerling L C and Michel J 2012 An electrically pumped germanium laser *Opt. Express* **20** 11316

[37] Di Cicco A, Polzoni G, Gunnella R, Trapananti A, Minicucci M, Rezvani S J, Catone D, Di Mario L, Pelli Cresi J S, Turchini S and Martelli F 2020 Broadband optical ultrafast reflectivity of Si, Ge and GaAs *Sci. Rep.* **10** 17363

[38] Karrock T and Gerken M 2015 Pressure sensor based on flexible photonic crystal membrane *Biomed. Opt. Express* **6** 4901

[39] Yang D, Tian H, Wu N, Yang Y and Ji Y 2013 Nanoscale torsion-free photonic crystal pressure sensor with ultra-high sensitivity based on side-coupled piston-type microcavity *Sensors Actuators A Phys.* **199** 30–6

[40] Lee C, Radhakrishnan R, Chen C-C, Li J, Thillaigovindan J and Balasubramanian N 2008 Design and Modeling of a Nanomechanical Sensor Using Silicon Photonic Crystals *J. Light. Technol.* **26** 839–46

[41] Li B and Lee C 2011 NEMS diaphragm sensors integrated with triple-nano-ring resonator *Sensors Actuators A Phys.* **172** 61–8

Accepted Manuscript

[42] Li B, Hsiao F-L and Lee C 2011 Computational Characterization of a Photonic Crystal Cantilever Sensor Using a Hexagonal Dual-Nanoring-Based Channel Drop Filter *IEEE Trans. Nanotechnol.* **10** 789–96

[43] John D. Joannopoulos, Steven G. Johnson, Joshua N. Winn R D M 2008 *Photonic crystals: Molding the flow of light* (New Jersey: Princeton University Press)

[44] Gamare K and Jain R B 2019 Performance analysis of 2D photonic crystal with line defect p 020011

[45] Kumar A and Kumar Goyal A 2024 Introductory Chapter: Photonic Crystal Technology – Introduction, Advantage, and Applications *Recent Adv. Trends Photonic Cryst. Technol.* 1–11

[46] Biswas U, Rakshit J K, Suthar B, Kumar D and Nayak C 2022 Modeling and analysis of all-optical pressure sensor using photonic crystal based micro ring resonator *Int. J. Numer. Model. Electron. Networks, Devices Fields* **35**

[47] Prasad P N 2004 *Nanophotonics* (Wiley)

[48] Pezzagna S, Brault J, Leroux M, Massies J and de Micheli M 2008 Refractive indices and elasto-optic coefficients of GaN studied by optical waveguiding *J. Appl. Phys.* **103**

[49] Elsayed H A, Sharma A, Alrowaili Z A and Taha T A 2021 Theoretical investigation of pressure sensing using a defect of polystyrene inside photonic crystals *Mater. Chem. Phys.* **270** 124853

[50] Huang M 2003 Stress effects on the performance of optical waveguides *Int. J. Solids Struct.* **40** 1615–32

[51] Goi A R, Syassen K and Cardona M 1990 Effect of pressure on the refractive index of Ge and GaAs *Phys. Rev. B* **41** 10104–10

[52] Cai J, Ishikawa Y and Wada K 2013 Strain induced bandgap and refractive index variation of silicon *Opt. Express* **21** 7162

[53] Servatkhah M and Pourmand R 2020 Optical properties of a two-dimensional GaAs quantum dot under strain and magnetic field *Eur. Phys. J. Plus* **135** 754

[54] Thamer A. Tabbakh, Raghad M. Aljohany H A and R M A 2021 Elastic, Optical, Transport, and Structural Properties of GaAs *Post-Transition Metals*

[55] T. Dharchana, A. Sivanantharaja S S 2017 Design of Pressure Sensor Using 2D Photonic Crystal *Adv. Nat.*

Accepted Manuscript

Appl. Sci. **11** 26–30

[56] Vijaya Shanthi K and Robinson S 2014 Two-dimensional photonic crystal based sensor for pressure sensing *Photonic Sensors* **4** 248–53

[57] R G P, L L D and rao kolli D 2018 Design of Two Dimensional Photonic Crystal Based Pressure Sensor Using Diaphragm *Int. Res. J. Eng. Technol.* 156–8

[58] Bahaddur I, Tejaswini M R, Kumar T.C. S, Sharan P and Srikanth P C 2019 2D Photonic Crystal Cantilever Resonator Pressure Sensor *2019 Workshop on Recent Advances in Photonics (WRAP)* (IEEE) pp 1–4

[59] Sreenivasulu T, Kolli V R, Anusree K, Yadunath T R, Badrinarayana T, Srinivas T, Hegde G and Mohan S 2015 Photonic crystal based force sensor on silicon microcantilever *2015 IEEE SENSORS* (IEEE) pp 1–4

[60] Yang Y, Yang D, Tian H and Ji Y 2013 Photonic crystal stress sensor with high sensitivity in double directions based on shoulder-coupled aslant nanocavity *Sensors Actuators A Phys.* **193** 149–54

[61] Li L, Li T, Song W, Zhang G and Li Y 2013 Design and Modeling of a Novel Two Dimensional Nano-Scaled Force Sensor Based On Silicon Photonic Crystal *ECS Meet. Abstr.* **MA2013-02** 2797–2797

[62] Tao S, Chen D, Wang J, Qiao J and Duan Y 2016 A high sensitivity pressure sensor based on two-dimensional photonic crystal *Photonic Sensors* **6** 137–42

[63] Olyaee S and Dehghani A A 2012 High resolution and wide dynamic range pressure sensor based on two-dimensional photonic crystal *Photonic Sensors* **2** 92–6

[64] Fu H Y, Wu C, Tse M L V., Zhang L, Cheng K-C D, Tam H Y, Guan B-O and Lu C 2010 High pressure sensor based on photonic crystal fiber for downhole application *Appl. Opt.* **49** 2639

[65] Okandan M, Olsson R, Baker M, Resnick P, Hill T A, Lackey C, Pearson S, Castaneda J, Trott W and Jones D 2009 High speed (GHZ), ultra-high pressure (GPA) sensor array fabricated in integrated CMOS+MEMS process *Proc. IEEE Int. Conf. Micro Electro Mech. Syst.* 845–7

Accepted Manuscript

The Journal of Supercomputing
https://doi.org/10.1007/s11227-024-06240-4

Probabilistic multiplicative unbalanced linguistic game using linguistic cloud model

Parul Chauhan[1] · **Anjana Gupta**[1]

Accepted: 15 May 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
In many game theoretic environments, the precise crisp values of payoffs are not always easy to gather as real-life decision-making problems are not inevitably precise and symmetrically distributed in nature. Hence, there is a need to explore matrix games having unbalanced probabilistic linguistic information. To fill this research gap, this study proposes a concept of probabilistic multiplicative unbalanced linguistic (PM-UL) game in which the players express their responses in terms of PM-UL term. This type of PM-UL game is a useful technique for multiple criteria analysis. Further, to thoroughly capture the uncertainty involved in payoffs of such games, they are converted to probabilistic unbalanced linguistic clouds (PULCs) that describe the uncertainties of qualitative payoffs by implementing the uncertain transformation between PM-UL payoffs and quantitative values. The PULCs are represented by three numerical characteristics, i.e. expectation, entropy and hyper-entropy. Hence, this model overcomes the inherent limitation of existing linguistic games where only the fuzziness of concepts is evaluated, by providing a PULC strategy and PULC value of the game. This procedure overcomes the loss of information occurring in the process of transformation of payoff values into quantitative ones. The defined methodology is implemented on a real-life working problem to demonstrate the effect and applicability in decision-making circumstances.

Keywords Probabilistic unbalanced linguistic term set · Normal cloud model · Two-person zero sum games · Probabilistic linguistic information

✉ Anjana Gupta
anjanagupta@dce.ac.in

[1] Department of Applied Mathematics, Delhi Technological University, Bawana Road, Rohini, Delhi, India

Published online: 01 June 2024

1 Introduction

Game theory has already proved its tremendous potential for conflict resolution problems in decision-making problems. Decision making is a social activity that is characterized by inherent uncertainty and ambiguous information communicated in day-to-day language. When a game theoretic approach is implemented to solve any decision-making problem, it is necessary to analyse which concept should be implemented to solve that particular problem. The results and methods of a problem are not always realistic and do not always match a real-life phenomenon. The ambiguity present in a concept make it difficult to assess a decision-making problem in a particular way. Hence, it is necessary to consider the uncertainty pertaining in a decision maker's (DM's) judgement to solve a problem underlying in a fuzzy environment. The development of fuzzy set theory [49] made it possible to deal with imprecise information explicitly. Several academics have looked into matrix games with payoffs that are fuzzy numbers [7, 8, 10, 14, 22, 25, 32] in order to reflect the uncertainty in the context of payoffs and goals. Also, the fuzzy programming approach plays a vital role in managing the group decision-making problems [35–38].

Although fuzzy games have provided phenomenal results in literature, these are still incapable of addressing games having qualitative information as the best expression in strategic forms occurs in the form of natural language only. Thus linguistic variables [50] play an important role in real-life decision making. Arfi [1, 2] introduced game theory based on fuzzy linguistic logic, which is a more straightforward and adaptable way to define payoffs in terms of linguistic labels. The utility of linguistic concepts implies a broad outlook in the current specialized literature on Computing with Words (CWW)-related methodologies. Also, a number of linguistic computational models based on balanced [17, 18] as well as unbalanced linguistic data [6, 9, 15, 28, 44, 52] are developed in existing literature to avoid information loss occurring in the process of CWW. For example, Herrera et al. [19] defined a computational model based on hierarchical structure to manage unbalanced linguistic information. Zhang et al. [54] revised the Herrera's model to define a simplified computational model to fuse multi-granular unbalanced linguistic terms. In fact, some authors have focused on developing linguistic computation models to managing multigranular unbalanced hesitant fuzzy linguistic information [53] in multi-attribute large-scale group decision making. Further, it is also necessary to model DMs' personalized individual semantics (PISs) in linguistic group decision making [55]. Zhang and Li's [51] work focuses on developing some PIS-based consistency control and consensus reaching models for this purpose. Furthermore, the DMs have extended the use of balanced as well as unbalanced linguistic information to deal with matrix games having such linguistic information [12, 13, 33, 34].

However, to deal with the ambiguity involved in the qualitative concepts, it is sometimes not possible to express the information with the help of a single particular term, or there are probabilities of occurrence of a qualitative term attached with them. For example, if $\mathbb{LT} = \{\text{Good}, \text{Very Good}\}$ is a linguistic term set, then it might be possible that the player expresses his linguistic information as "50%

chances of being good", then we might say that the linguistic term “Good" has a probability attached as “Good (0.5)". For expressing such type of information, there is a need to use probabilistic linguistic term sets (PLTS) introduced in [31]. The PLTS enables the players or DMs to evaluate using numerous linguistic terms rather than a single term, more efficiently by giving each term a weight through probability. Moreover, a PLTS may be transformed into linguistic distribution so that each and every term of the original linguistic set is present in the final evaluation by the DMs [45]. A lot of research has been carried out based on PLTSs [3, 20, 46–48] as these types of sets allow the DMs to use more than one linguistic term with associated probabilities. Mi et al. [30] put forward this extensive and comprehensive way of expressing information in matrix games taking PLTS as input of matrix game. However, this PLTS consists of symmetrically distributed linguistic terms only. As it is well known that the real-life decision-making problems are not just based on symmetric data sets. Rather, a practical decision-making problem considers terms that are irregularly distributed. In order to deal with unbalanced linguistic information, Han et al. [16] established probabilistic unbalanced linguistic term set (PULTS), in which the authors proposed a computational model based on copula and co-copula. However, an additive scale set is still used to represent the data, therefore the term “unbalanced” does not retain its literal meaning. This is the first research gap in the literature that is addressed in the present study.

To handle unbalanced linguistic probabilistic information, Malhotra and Gupta [27] introduced a probabilistic multiplicative ULTS (PM-ULTS) that takes into account the irregular distribution of probabilistic linguistic terms and is sufficient enough to deal with all types of ULTSs. Further they have solved matrix games that have payoffs in terms of PM-ULTS. By the concept “PM-UL payoffs”, it is meant that the players of the game express their responses in the form of linguistic terms that are distributed asymmetrically and are multiplicative in their distribution with some attached probabilities. Moreover, the set consisting of such types of linguistic terms is referred to as probabilistic multiplicative unbalanced linguistic term set. However, despite such study of PM-UL game, the methodology defined by [27] also has a limitation that it handles only one kind of uncertainty, i.e. fuzziness. The two most significant uncertainties in natural language, however, are randomness and fuzziness that have received a lot of attention in the field of artificial intelligence. It is typically exceedingly challenging to ascertain precise information or define the bounds of extensions if we only take fuzziness into account. For instance, there is no specific value that can be used to describe the concept’s distinct border when we define the linguistic term “old". This is the second research gap addressed in the current study. In general, it is always possible for a linguistic variable to take on distinct meanings for different people, and even for the same person, to have different interpretations in different circumstances due to the abstractness and generality of notions. As a result, some notions frequently include randomness as well. Many linguistic terms have both randomness and fuzziness at the same time, despite the fact that they are of very different kinds. For instance, the linguistic terms “about", “maybe", “probably", etc., although represent random thoughts, also exhibit the fuzziness due to the vague boundary of the extension. In the light of this, it seems more practical for a formal

representation model of concepts to simultaneously account for both randomness and the fuzziness. Li et al. [21] presented a cloud model that can synthetically express the concept of randomness and fuzziness of decision information at the same time to formally characterize the underlying relationship between randomness and fuzziness and effectively integrate both uncertainties. The cloud models use EX (expectation), EN (entropy), and HE (hyperentropy) to illustrate a linguistic concept. Since the development of cloud models, numerous academics have dedicated their efforts to advance cloud-specific studies [23, 24, 26, 29, 39–42]. Although existing cloud models are believed to be able to preserve the original information, they are still unable to accurately represent labels that are asymmetrically distributed. To solve this problem, Wang et al. [43] proposed an Asymmetric Trapezoidal Clouds (ATCs) model as a solution to this problem.The EN and HE of the ATCs differ on the left and right sides in this model, and the authors expressed the expectation as an interval rather than a specific number. To address multiplicative unbalanced linguistic information, Chauhan et al. [11] have developed an accurate and innovative cloud decision support model. Here, in the present paper, we solve two-person matrix games having payoffs in the form of PM-ULTS by dealing with both randomness and fuzziness of the linguistic concepts. Further, the operational laws used to aggregate the payoffs in the paper [27] are based on probabilistic linguistic terms, while in present paper, well-defined cloud aggregation operators are used that are already well established in literature and do not follow any complex representation. This is the third research gap to be addressed in the current paper.

To fill all the discussed research gaps, the following contributions are offered in this paper:

- The matrix games having payoffs in the form of PM-ULTS are considered and solved using the defined methodology. These PM-ULTSs are able to handle all types of regular as well as irregular distribution of the linguistic concepts. Moreover, the defined methodology is able to deal with the most important uncertainties, i.e. randomness and fuzziness underlying in a qualitative information.
- An unbalanced linguistic cloud model is used to transform the probabilistic multiplicative data into its numerical form without making any significant loss of information. The probabilistic multiplicative unbalanced linguistic variables that are used as payoffs of a game problem are transformed into probabilistic unbalanced linguistic clouds (PULCs), i.e. the clouds generated from the probabilistic unbalanced linguistic terms, to effectively capture the uncertainty of the probabilistic linguistic payoff.
- The dispersion of cloud drops is measured by standard deviation S.D. defined in the paper that suggests the relative stability distribution of the linguistic term. Also, the score function is defined to compare different PULCs.
- To address the practical issues, the defined approach is then applied to address matrix games having probabilistic multiplicative unbalanced linguistic payoffs. The values of the players are defined in terms of PULCs. Also, the proposed method provides more accurate results by providing the values and strategies of the game in terms of expectation, entropy and hyperentropy. The expectation is the most representative and typical sample of the qualitative concept,

while entropy represents the randomness and fuzziness extent of the cloud drop. Hyperentropy reflects the uncertainty degree of entropy.

The paper is organized in the following manner. Section 2 reviews the basic concepts related to PM-ULTS, unbalanced linguistic cloud model and two-person matrix games that are required in the present paper. Section 3 provides the transformation of probabilistic multiplicative unbalanced linguistic variables to PULCs and operational laws. In Sect. 4, the application of PULC model in two-person constant sum matrix games is illustrated by first defining a PM-ULC game (PM-ULCG) along with the methodology for solving PM-ULCG followed by a numerical problem. Some analysis of results obtained and comparison with other defined methodologies, along with discussion, is provided in Sect. 5. Section 6 concludes the paper along with future scope.

2 Preliminaries

This section reviews some basic concepts that are required in this paper. The PM-ULTS defined in Sect. 2.1 consists of unbalanced linguistic terms having probabilities associated with them. Section 2.2 discusses the unbalanced linguistic cloud model based on which the payoffs are aggregated in this paper. In Sect. 2.3, the fundamentals of two-person matrix games are discussed.

2.1 Probabilistic multiplicative unbalanced linguistic term set (PM-ULTS)

Definition 1 [27] Let $\mathbb{LS} = \{t_{a^i} | i = -n, \ldots, 0, \ldots, m\}$ be a ULTS, where n and m may or may not be equal, $a > 1$. Then, a probabilistic multiplicative ULTS is defined as

$$\mathbb{LS}(p) = \{t_{a^i}{}^k(p^k) | k = 1, 2, \ldots, \#\mathbb{LS}(p)\}$$

where $t_{a^i}{}^k \in \mathbb{LS}$, $p^k \geq 0$, $\sum_{k=1}^{\#\mathbb{LS}(p)} p^k$ may or may be not be equal to 1 and $\#\mathbb{LS}(p)$ is the cardinality of all the given distinct multiplicative unbalanced linguistic terms in $\mathbb{LS}(p)$. The term $t_{a^i}{}^k(p^k)$ is interpreted as unbalanced linguistic term $t_{a^i}{}^k$ associated with the probability p^k.

Note If $\sum_{k=1}^{\#\mathbb{LS}(p)} p^k < 1$ or $\sum_{k=1}^{\#\mathbb{LS}(p)} p^k > 1$, then the probability information within the set is incomplete and we normalize the incomplete PM-ULTS to transform the set into complete PM-ULTS using the following process.

1. Normalize the probability. If $\sum_{k=1}^{\#\mathbb{LS}(p)} p^k \neq 1$, then the given PM-ULTS is defined as $\widetilde{\mathbb{LS}}(p) = \{t_{a^i}{}^k(\bar{p}^k) | \sum_{k=1}^{\#\mathbb{LS}(p)} \bar{p}^k = 1\}$, where $\bar{p} = \frac{p^k}{\sum_{k=1}^{\#\mathbb{LS}(p)} p^k}$, $t_{a^i} \in \mathbb{LS}$, $p^k \geq 0$, $k = 1, \ldots, \#\mathbb{LS}(p)$;

2. Normalize the granularity. If $\mathbb{LS}_1(p)$ and $\mathbb{LS}_2(p)$ are two PM-ULTS. If $\#\mathbb{LS}_1(p) < \#\mathbb{LS}_2(p)$, add $\#\mathbb{LS}_2(p) - \#\mathbb{LS}_1(p)$ multiplicative unbalanced linguistic terms to $\mathbb{LS}_1(p)$

Definition 2 [27] Let $\mathbb{LS}(p) = \{t_{a^i}{}^k(\bar{p}^k)|k = 1, \ldots, \#\mathbb{LS}(p)\}$ be a normalized PM-ULTS. The negation of $\mathbb{LS}(p)$ is defined as $Neg(\mathbb{LS}(p)) = \{neg(t_{a^i}{}^k)(\bar{p}^k)|k = 1, \ldots, \#\mathbb{LS}(p)\}$.

Definition 3 [27] Let $\mathbb{LS}_1(p_1) = \{t_{a^i}{}^k(p^k)|k = 1, 2, \ldots, \#\mathbb{LS}_1(p_1)\}$ and $\mathbb{LS}_2(p_2) = \{t_{a^j}{}^k(p^k)|k = 1, 2, \ldots, \#\mathbb{LS}_2(p_2)\}$, $i,j = -n, \ldots, m$ be any two ordered PM-ULTS. Then, the operational laws are given as follows:

1. $\mathbb{LS}_1(p_1) \oplus \mathbb{LS}_2(p_2) = \{t^k_{a^i \oplus a^j}(\bar{p}^k_1 \oplus \bar{p}^k_2)|k = 1, 2, \ldots, \#\mathbb{LS}_1(p_1)\}$
2. $\mathbb{LS}_1(p_1) \otimes \mathbb{LS}_2(p_2) = \{t^k_{a^i \otimes a^j}(\bar{p}^k_1 \otimes \bar{p}^k_2)|k = 1, 2, \ldots, \#\mathbb{LS}_1(p_1)\}$
3. Let $\lambda \in (0, 1)$ be any arbitrary scalar, $\lambda \odot \mathbb{LS}(p) = \{t^k{}_{\lambda \odot a^i}(\bar{p}^k)|k = 1, 2, \ldots \mathbb{LS}(p)\}$
4. $(\mathbb{LS}(p))^\lambda = \{t^k_{a^{\lambda i}}(\bar{p}^k)|k = 1, 2, \ldots \#\mathbb{LS}(p)\}$

2.2 Unbalanced linguistic cloud model

Definition 4 [11] Let $\mathbb{LS} = \{t_{a^i}|i = -n, \ldots, 0, \ldots, m\}$, $a > 1$ (n and m may or may not be equal) be any finite ordered multiplicative unbalanced LT set. The linguistic scale function (LSF) converting a linguistic term t_{a^i} to a certain number δ_{a^i}, $(i = -n, \ldots, 0, \ldots, m)$ can be expressed as follows:

$$\widetilde{\mathfrak{g}} : t_{a^i} \to \delta_{a^i} \, (i = -n, \ldots, 0, \ldots, m)$$

where

$$\delta_{a^i} = \begin{cases} \left(\frac{b^n - b^{-i}}{2b^n - 2}\right)\frac{1}{\log a}, & -n \le i \le 0 \\ \left(\frac{b^n + b^i - 2}{2b^n - 2}\right)\frac{1}{\log a}, & 0 < i \le n \end{cases} \tag{1}$$

Here, $\delta_{a^i} \in [0, 1/\log a] \subseteq [0, 1]$ is a numeric value. The value of the parameter b can be obtained through experiments such as $b \in [1.36, 1.4]$ [4]. In this paper, we take $b = 1.4$ without any loss of generality.

The next definition introduces the process of generating cloud droplets based on the rational transformation method of unbalanced linguistic variables into clouds.

Definition 5 [11] Let $t_{a^i} \in \mathbb{LS}$ be any arbitrary unbalanced linguistic term corresponding to the cloud drop $(EX_{a^i}, EN_{a^i}, HE_{a^i})$. Then, the working cloud generation procedure is given as follows:

1. Calculate δ_{a^i} using Eq. 1.
2. Calculate EX_{a^i}.

$$EX_{a^i} = X_{\min} + \delta_{a^i}(X_{\max} - X_{\min})$$

3. Calculate EN_{a^i}.

$$3EN'_{a^i} = \max(X_{\max} - EX_{a^i}, EX_{a^i} - X_{\min})$$

Because of $EN' \sim N(EN, HE^2)$, EN_{a^i} may be considered the mean value of EN'_{a^i} corresponding to the clouds in consideration and their adjacent clouds.

$$EN_{a^{-i}} = EN_{a^i} = \frac{EN'_{a^{i-1}} + EN'_{a^i} + EN'_{a^{i+1}}}{3\log a} \text{ for } |i| = 0, 1, \ldots, n-1, \tag{2}$$

and for $|i| = n$,

$$EN_{a^{-n}} = EN_{a^n} = \frac{EN'_{a^n} + EN'_{a^{n-1}}}{2\log a} \tag{3}$$

These calculations may be rewritten as the following formulas:

$$EN'_{a^i} = \begin{cases} \frac{(1-\delta_{a^i})(X_{\max}-X_{\min})}{3\log a}, & -n \le i \le 0 \\ \frac{\delta_{a^i}(X_{\max}-X_{\min})}{3\log a}, & 0 < i \le n \end{cases} \tag{4}$$

$$EN_{a^{-i}} = EN_{a^i} = \begin{cases} \frac{(\delta_{a|i-1|}+\delta_{a|i|}+\delta_{a|i+1|})(X_{\max}-X_{\min})}{9\log a}, & 0 < |i| < n-1 \\ \frac{(\delta_{a|i|}+\delta_{a|i-1|})(X_{\max}-X_{\min})}{6\log a}, & |i| = n \\ \frac{(\delta_{a^i}+2\delta_{a^{i+1}})(X_{\max}-X_{\min})}{9\log a}, & i = 0 \end{cases} \tag{5}$$

4. Calculate HE_{a^i}. Due to $EN'_{a^i} \sim N(EN_{a^i}, {HE_{a^i}}^2)$, the hyperentropy satisfies the 3σ principle. In particular,

$$HE_{a^{-i}} = HE_{a^i} = \frac{EN'^{+} - EN_{a^i}}{3\log a} \tag{6}$$

where, $EN'^{+} = \sup_k\{EN'_k\}$.

Note It should be noted that the point-based cloud-generating procedure described above can be used in scenario $n \neq m$ if dummy variables are added to the multiplicative unbalanced linguistic term set in a way that $n = m$.

2.3 Matrix games

Definition 6 [5] A crisp two-player zero-sum matrix game is represented as a triplet (S^r, S^s, A), where S^r is the set of 'r' possible strategies of Player I, S^s is the set of 's' possible strategies of Player II, $[A]_{r\times s}$ is the payoff matrix of order of Player I corresponding to Player II, and $neg(A)$ refers to the matrix consisting of payoffs of Player II.

Definition 7 [5] A matrix game having $A = (a_{ij})_{r\times s}$ as the matrix of payoffs has the lower value

$$v^- = \max_{i=1,\ldots,r} \min_{j=1,\ldots,s} a_{ij}$$

and the upper value

$$v^+ = \min_{j=1,\ldots,s} \max_{i=1,\ldots,r} a_{ij}$$

The upper value v^+ is the guaranteed maximum that Player II may lose, while the lower value v^- is the guaranteed minimum that Player I is guaranteed to get. If $v^+ = v^-$, we say that the game has a value of v.

Definition 8 [5] If no pure strategies exist, a mixed strategy is a vector where $\Phi = (\phi_1, \phi_2, \ldots, \phi_r)$ for Player I and $\Psi = (\psi_1, \psi_2, \ldots, \psi_s)$ for Player II, where

$$\phi_i \geq 0, \sum_{i=1}^{r} \phi_i = 1 \quad \& \quad \psi_j \geq 0, \sum_{j=1}^{s} \psi_j = 1.$$

Player I's probability of utilizing row 'i' and Player II's probability of using column 'j' are each represented by ϕ_i and ψ_j. We represent the collection of mixed strategies having 'k' components as

$$S^k = \left\{ (z_1, z_2, \ldots, z_k) | \, z_r \geq 0, r = 1, 2, \ldots, k, \sum_{r=1}^{k} z_r = 1 \right\}.$$

Given a choice of mixed strategy $\Phi = (\phi_1, \phi_2, \ldots, \phi_r)$ for Player I and $\Psi = (\psi_1, \psi_2, \ldots, \psi_s)$ for Player II, selected autonomously, the expected payoff of Player I is

$$\begin{aligned} E(\Phi, \Psi) &= \sum_{i=1}^{r} \sum_{j=1}^{s} \phi_i a_{ij} \psi_j \\ &= \Phi A \Psi^T \end{aligned}$$

3 Transformation of PM-UL variables to PULCs and operational laws

In this section, we present how to transform a PM-UL variable to its corresponding PULC, and a numerical example is shown. Further, the operational laws and score function of PULCs are defined.

3.1 Method for transformation of probabilistic unbalanced linguistic variables to probabilistic unbalanced linguistic clouds (PULCs)

Definition 9 Let $\mathbb{LS} = \{t_{a^i} | i = -n, \ldots, 0, \ldots, m\}$ be a ULTS where n and m may or may not be equal and $a > 1$. Suppose p_i be any probability associated with the unbalanced linguistic term t_{a^i} such that $\mathbb{LS}(p) = \{t_{a^i}(p_i) | i = -n, \ldots, m\}$ denotes the linguistic set representing linguistic terms having attached probabilities.

Consider a valid universe $X = [X_{\min}, X_{\max}]$. Now to represent the linguistic terms of the set $\mathbb{LS}(p)$, the procedure is as given below.

- First transform the linguistic terms into their corresponding cloud using the method given in Definition 4 and 5, i.e.

* Find the value of the LSF δ_{a^i}, $(i = -n, \ldots, 0, \ldots, m)$
* Calculate EX_{a^i} as $EX_{a^i} = X_{\min} + \delta_{a^i}(X_{\max} - X_{\min})$.
* Compute EN_{a^i}.
* Calculate HE_{a^i}.

Note. The above-mentioned procedure is used in the scenario $n = m$. In case $n \neq m$, then we add a dummy linguistic term with probability $\}0'$ in $\mathbb{LS}(p)$ so that $n = m$, where the linguistic term is added on the side of $\min(m, n)$.

- Now consider the attached probabilities with the linguistic terms by multiplying these probabilities with the above-obtained cloud. Hence, a cloud corresponding to a linguistic term $t_{a^i}(p_i)$ is given as $\mathbb{Y}_i(p_i EX_{a^i}, p_i EN_{a^i}, p_i HE_{a^i})$.

$$\therefore t_{a^i}(p_i) \longrightarrow (p_i EX_{a^i}, p_i EN_{a^i}, p_i HE_{a^i})$$

Example 1 Given the universe $[X_{\min}, X_{\max}] = [0, 7]$, let $\mathbb{LS} = \{t_{2^{-2}} : \text{Terrible}, t_{2^{-1}} : \text{Poor}, t_{2^0} : \text{Middle}, t_{2^1} : \text{Satisfactory}, t_{2^2} : \text{Good}, t_{2^3} : \text{Excellent}\}$. As $n \neq m$, we add a dummy linguistic term on the left hand side of $\mathbb{LS}$ with probability $\}0'$. Consider $\mathbb{LS}(p) = \{t_{2^{-3}}(0), t_{2^{-2}}(0), t_{2^{-1}}(0.1), t_{2^0}(0.4), t_{2^1}(0), t_{2^2}(0.2), t_{2^3}(0.3)\}$. By following the steps mentioned above, the cloud is generated as follows. Setting $b = 1.4$.

1. First we calculate δ_{a^i} for each t_{a^i}. $\delta_{2^{-3}} = 0 \quad \delta_{2^{-2}} = 0.7467 \quad \delta_{2^{-1}} = 1.2801$ $\delta_{2^0} = 1.6611 \quad \delta_{2^1} = 2.0421 \quad \delta_{2^2} = 2.5755 \quad \delta_{2^3} = 3.3222$
2. $EX_{2^{-3}} = 0 \quad EX_{2^{-2}} = 5.22718 \quad EX_{2^{-1}} = 8.9607 \quad EX_{2^0} = 11.6277 \quad EX_{2^1} = 14.2947$ $EX_{2^2} = 18.0285 \quad EX_{2^3} = 23.2557$
3. $EN'_{2^{-3}} = 7.7519 \quad EN'_{2^{-2}} = 1.9632 \quad EN'_{2^{-1}} = -2.1715 \quad EN'_{2^0} = -5.125$ $EN'_{2^1} = 15.8308 \quad EN'_{2^2} = 19.9652 \quad EN'_{2^3} = 25.7538 \therefore \max EN'_{a^i} = 25.7538$ Next, $EN_{2^{-3}} = EN_{2^3} = 22.8595 \quad EN_{2^{-2}} = EN_{2^2} = 20.5164 \quad EN_{2^{-1}} = EN_{2^1} = 16.2241$ $EN_{2^0} = 14.8457$
4. $HE_{2^{-3}} = HE_{2^3} = 20.9258 \quad HE_{2^{-2}} = HE_{2^2} = 23.9762 \quad HE_{2^{-1}} = HE_{2^1} = 26.7236$ $HE_{2^0} = 21.944$

Hence, the probabilistic multiplicative unbalanced clouds corresponding to each linguistic term of the set $\mathbb{LS}(p)$ are given as: $t_{2^{-3}}(0) \longrightarrow (0, 0, 0) \quad t_{2^{-2}}(0) \longrightarrow (0, 0, 0)$ $t_{2^{-1}}(0.1) \longrightarrow (0.89607, 1.6224, 2.6723) \quad t_{2^0}(0.4) \longrightarrow (4.651, 5.9382, 8.777) \quad t_{2^1}(0) \longrightarrow$ $t_{2^1}(0) \longrightarrow (0, 0, 0) \quad t_{2^2}(0.2) \longrightarrow (3.6057, 4.1032, 4.7952) \quad t_{2^3}(0.3) \longrightarrow (6.9767,$ $6.8578, 6.2777)$ These probabilistic multiplicative unbalanced linguistic clouds are represented in Fig. 1.

In the graph, the dispersing extent of cloud drops is depicted by entropy and the hyperentropy EN and HE, respectively. It is also the measurement of fuzziness, representing the scope of the universe that can be accepted by the concept.

3.2 Operations and score functions of probabilistic unbalanced linguistic clouds

In this section, we define some operational laws for probabilistic unbalanced linguistic clouds based on the operators defined for normal clouds.

Definition 10 Given the PULCs $\mathbb{Y}_i = (p_i EX_{a^i}, p_i EN_{a^i}, p_i HE_{a^i})$, the basic operations are developed as follows:

1. $\sum_{i=-n}^{m} (p_i EX_{a^i}, p_i EN_{a^i}, p_i HE_{a^i}) = \left(\sum_{i=-n}^{m} p_i EX_{a^i}, \sqrt{\sum_{i=-n}^{m} \left(p_i EN_{a^i}\right)^2}, \sqrt{\sum_{i=-n}^{m} \left(p_i HE_{a^i}\right)^2} \right)$
2. $\lambda(p_i EX_{a^i}, p_i EN_{a^i}, p_i HE_{a^i}) = (\lambda p_i EX_{a^i}, \lambda p_i EN_{a^i}, \lambda p_i HE_{a^i})$, where λ is any real scalar.
3. Let $\mathbb{Y}_i(p_i EX_{a^i}, p_i EN_{a^i}, p_i HE_{a^i})$, $i = -n, \ldots, m$ be a collection of PULCs with associated weights $\omega_i \in [0, 1]$ and $\sum_{i=-n}^{m} \omega_i = 1$. Then, their aggregated value with use of the weighted average (WA) operator is still a cloud and is denoted as follows:

$$WA(Y_{-n}, \ldots, Y_m) = \left(\sum_{i=-n}^{m} \omega_i p_i EX_{a^i}, \sqrt{\sum_{i=-n}^{m} \omega_i \left(p_i EN_{a^i}\right)^2}, \sqrt{\sum_{i=-n}^{m} \omega_i \left(p_i HE_{a^i}\right)^2} \right) \quad (7)$$

The score function is a crucial tool for obtaining the ranking results during the decision-making process. For the normal clouds, a number of score functions have previously been defined. Here, we define the score function for PULCs. Before

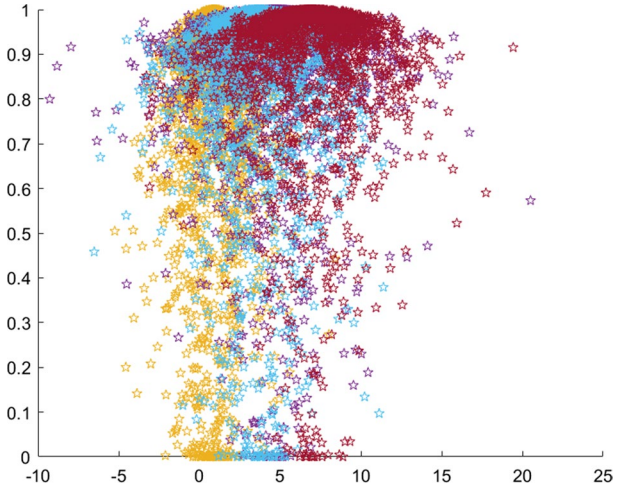


Fig. 1 Representation of probabilistic multiplicative unbalanced linguistic clouds with respect to linguistic terms having attached probabilities based on 1000 droplets

presenting the new function, we first introduce an important concept, the variance of a normal cloud, which is developed by Li et al. [30].

Definition 11 Let $(p_iEX_{a^i}, p_iEN_{a^i}, p_iHE_{a^i})$ be a PULC with $n+m+1$ generated cloud drops, and the variance of these cloud drops can be computed as

$$V = \left(p_iEN_{a^i}\right)^2 + \left(p_iHE_{a^i}\right)^2 \tag{8}$$

Hence, the standard deviation follows $S.D. = \sqrt{V}$.

The standard deviation ($S.D.$), which can be used to express the ambiguity and fluctuation of a quantitative notion, is used to measure the dispersion of cloud drops. A low standard deviation indicates that the distribution of the linguistic term is reasonably steady. A high standard deviation suggests stability even when the linguistic concept has a high degree of ambiguity and volatility. In this case, the relationship between EX and $S.D$ can be seen as the expected utility and utility dispersion of the utility theory. Therefore, the choice for an option (alternative) can be expressed as a linear trade-off between expected utility and utility dispersion. In response to this approach, we construct a score formula below to compare PULCs.

Definition 12 Given a PULC $(p_iEX_{a^i}, p_iEN_{a^i}, p_iHE_{a^i})$, then a score function of PULC is defined as

$$\hat{S} = p_iEX_{a^i} - \gamma S.D. \tag{9}$$

where $S.D.$ is the standard deviation of a PULC, $\gamma \in [0, 1]$ refers to a coefficient to capture the attitude towards the standard deviation. The value of γ can be chosen based on actual decision-making circumstances. $\gamma = 0$ signifies that the DMs do not

care about the uncertainty and fluctuation of the assessment information and simply take into account the average level, which is the expectation *EX*. The greater the value of γ, the greater the weight given to *S.D*.

The score function $\hat{S}$ is monotonically increasing w.r.t. *EX* while monotonically decreasing w.r.t. *EN* and *HE*. On the basis of score function, two PULCs may be compared in the following manner.

Definition 13 Let $\mathbb{Y}_1$ and $\mathbb{Y}_2$ be two PULCs, and $\hat{S}_1$ and $\hat{S}_2$ denote their score functions, respectively, then

- If $\hat{S}_1 > \hat{S}_2$, then we prefer $\mathbb{Y}_1$ over $\mathbb{Y}_2$ and denote it as $\mathbb{Y}_1 > \mathbb{Y}_2$.
- $\hat{S}_1 = \hat{S}_2$ represents that both the PULCs are equally preferred and are denoted as $\mathbb{Y}_1 = \mathbb{Y}_2$.
- $\max\{\mathbb{Y}_i | i = -n, \ldots, m\} = \mathbb{Y}_{\max}$ such that $\hat{S}(\mathbb{Y}_{\max}) \geq \hat{S}(\mathbb{Y}_i)\ \forall\, i$.
- $\min\{\mathbb{Y}_i | i = -n, \ldots, m\} = \mathbb{Y}_{\min}$ such that $\hat{S}(\mathbb{Y}_{\min}) \leq \hat{S}(\mathbb{Y}_i)\ \forall\, i$.

Example 2 Consider the problem given in Example 1, the set consisting of linguistic terms with attached probabilities is $\mathbb{LS}(p) = \{t_{2^{-3}}(0),\ t_{2^{-2}}(0),\ t_{2^{-1}}(0.1),\ t_{2^0}(0.4), t_{2^1}(0), t_{2^2}(0.2),\ t_{2^3}(0.3)\}$. The PULCs obtained from these probabilistic unbalanced linguistic terms are

$$\mathbb{Y}_1(0,0,0) \quad \mathbb{Y}_2(0,0,0) \quad \mathbb{Y}_3(0.89607, 1.6224, 2.6723) \quad \mathbb{Y}_4(4.651, 5.9382, 8.777)$$

$$\mathbb{Y}_5(0,0,0) \quad \mathbb{Y}_6(3.6057, 4.1032, 4.7952) \quad \mathbb{Y}_7(6.9767, 6.8578, 6.2777)$$

Using the definition of score function, we find scores of these PULCs for some random values of γ and provide their rankings in Table 1.

Now, from the aspect of information representation form, the PM-UL information is a flexible tool for CWW. Also, the unbalanced linguistic cloud model has provided wonderful results to deal with uncertainty of a linguistic concept. But in the framework of game theory, it is still not investigated how to solve a game problem whose payoffs are of the form PM-UL by capturing the randomness and fuzziness involved in the payoffs. In this regard, Sect. 4 presents the definition of a PM-UL game (PM-ULG) and a methodology is given to solve such type of game by converting its payoffs to PULCs and treating it as a PM-UL cloud game (PM-ULCG).

Table 1 PULC rankings based on scores for different values of γ

Value of γ	$\hat{S}(\mathbb{Y}_1)$	$\hat{S}(\mathbb{Y}_2)$	$\hat{S}(\mathbb{Y}_3)$	$\hat{S}(\mathbb{Y}_4)$	$\hat{S}(\mathbb{Y}_5)$	$\hat{S}(\mathbb{Y}_6)$	$\hat{S}(\mathbb{Y}_7)$	Score ranking	PULC ranking
0	0	0	0.896	4.651	0	3.605	6.277	$\hat{S}(\mathbb{Y}_7) > \hat{S}(\mathbb{Y}_4) > \hat{S}(\mathbb{Y}_6) >$ $\hat{S}(\mathbb{Y}_3) > \hat{S}(\mathbb{Y}_1) = \hat{S}(\mathbb{Y}_2) = \hat{S}(\mathbb{Y}_5)$	$\mathbb{Y}_7 > \mathbb{Y}_4 > \mathbb{Y}_6 >$ $\mathbb{Y}_3 > \mathbb{Y}_1 = \mathbb{Y}_2 = \mathbb{Y}_5$
0.1	0	0	0.583	3.591	0	2.974	6.046	$\hat{S}(\mathbb{Y}_7) > \hat{S}(\mathbb{Y}_4) > \hat{S}(\mathbb{Y}_6) >$ $\hat{S}(\mathbb{Y}_3) > \hat{S}(\mathbb{Y}_1) = \hat{S}(\mathbb{Y}_2) = \hat{S}(\mathbb{Y}_5)$	$\mathbb{Y}_7 > \mathbb{Y}_4 > \mathbb{Y}_6 >$ $\mathbb{Y}_3 > \mathbb{Y}_1 = \mathbb{Y}_2 = \mathbb{Y}_5$
0.2	0	0	0.27	2.531	0	2.343	5.117	$\hat{S}(\mathbb{Y}_7) > \hat{S}(\mathbb{Y}_4) > \hat{S}(\mathbb{Y}_6) >$ $\hat{S}(\mathbb{Y}_3) > \hat{S}(\mathbb{Y}_1) = \hat{S}(\mathbb{Y}_2) = \hat{S}(\mathbb{Y}_5)$	$\mathbb{Y}_7 > \mathbb{Y}_4 > \mathbb{Y}_6 >$ $\mathbb{Y}_3 > \mathbb{Y}_1 = \mathbb{Y}_2 = \mathbb{Y}_5$
0.3	0	0	− 0.041	1.471	0	1.712	4.187	$\hat{S}(\mathbb{Y}_7) > \hat{S}(\mathbb{Y}_6) > \hat{S}(\mathbb{Y}_4) >$ $\hat{S}(\mathbb{Y}_1) = \hat{S}(\mathbb{Y}_2) = \hat{S}(\mathbb{Y}_5) > \hat{S}(\mathbb{Y}_3)$	$\mathbb{Y}_7 > \mathbb{Y}_6 > \mathbb{Y}_4 >$ $\mathbb{Y}_1 = \mathbb{Y}_2 = \mathbb{Y}_5 > \mathbb{Y}_3$
0.4	0	0	− 0.354	0.412	0	1.081	3.257	$\hat{S}(\mathbb{Y}_7) > \hat{S}(\mathbb{Y}_6) > \hat{S}(\mathbb{Y}_4) >$ $\hat{S}(\mathbb{Y}_1) = \hat{S}(\mathbb{Y}_2) = \hat{S}(\mathbb{Y}_5) > \hat{S}(\mathbb{Y}_3)$	$\mathbb{Y}_7 > \mathbb{Y}_6 > \mathbb{Y}_4 >$ $\mathbb{Y}_1 = \mathbb{Y}_2 = \mathbb{Y}_5 > \mathbb{Y}_3$
0.5	0	0	− 0.667	− 0.647	0	0.45	2.328	$\hat{S}(\mathbb{Y}_7) > \hat{S}(\mathbb{Y}_6) > \hat{S}(\mathbb{Y}_1) =$ $\hat{S}(\mathbb{Y}_2) = \hat{S}(\mathbb{Y}_5) > \hat{S}(\mathbb{Y}_4) > \hat{S}(\mathbb{Y}_3)$	$\mathbb{Y}_7 > \mathbb{Y}_6 > \mathbb{Y}_1 =$ $\mathbb{Y}_2 = \mathbb{Y}_5 > \mathbb{Y}_4 > \mathbb{Y}_3$
0.6	0	0	− 0.979	− 1.707	0	− 0.18	1.398	$\hat{S}(\mathbb{Y}_7) > \hat{S}(\mathbb{Y}_1) = \hat{S}(\mathbb{Y}_2) =$ $\hat{S}(\mathbb{Y}_5) > \hat{S}(\mathbb{Y}_6) > \hat{S}(\mathbb{Y}_3) > \hat{S}(\mathbb{Y}_4)$	$\mathbb{Y}_7 > \mathbb{Y}_1 = \mathbb{Y}_2 =$ $\mathbb{Y}_5 > \mathbb{Y}_6 > \mathbb{Y}_3 > \mathbb{Y}_4$
0.7	0	0	− 1.292	− 2.766	0	− 0.812	0.468	$\hat{S}(\mathbb{Y}_7) > \hat{S}(\mathbb{Y}_1) = \hat{S}(\mathbb{Y}_2) =$ $\hat{S}(\mathbb{Y}_5) > \hat{S}(\mathbb{Y}_6) > \hat{S}(\mathbb{Y}_3) > \hat{S}(\mathbb{Y}_4)$	$\mathbb{Y}_7 > \mathbb{Y}_1 = \mathbb{Y}_2 =$ $\mathbb{Y}_5 > \mathbb{Y}_6 > \mathbb{Y}_3 > \mathbb{Y}_4$
0.8	0	0	− 1.604	− 3.826	0	− 1.443	− 0.461	$\hat{S}(\mathbb{Y}_1) = \hat{S}(\mathbb{Y}_2) = \hat{S}(\mathbb{Y}_5) >$ $\hat{S}(\mathbb{Y}_7) > \hat{S}(\mathbb{Y}_6) > \hat{S}(\mathbb{Y}_3) > \hat{S}(\mathbb{Y}_4)$	$\mathbb{Y}_1 = \mathbb{Y}_2 = \mathbb{Y}_5 >$ $\mathbb{Y}_7 > \mathbb{Y}_6 > \mathbb{Y}_3 > \mathbb{Y}_4$
0.9	0	0	− 1.917	− 4.886	0	− 2.074	− 1.39	$\hat{S}(\mathbb{Y}_1) = \hat{S}(\mathbb{Y}_2) = \hat{S}(\mathbb{Y}_5) >$ $\hat{S}(\mathbb{Y}_7) > \hat{S}(\mathbb{Y}_3) > \hat{S}(\mathbb{Y}_6) > \hat{S}(\mathbb{Y}_4)$	$\mathbb{Y}_1 = \mathbb{Y}_2 = \mathbb{Y}_5 >$ $\mathbb{Y}_7 > \mathbb{Y}_3 > \mathbb{Y}_6 > \mathbb{Y}_4$
1	0	0	− 2.239	− 5.946	0	− 2.705	− 2.32	$\hat{S}(\mathbb{Y}_1) = \hat{S}(\mathbb{Y}_2) = \hat{S}(\mathbb{Y}_5) >$ $\hat{S}(\mathbb{Y}_3) > \hat{S}(\mathbb{Y}_7) > \hat{S}(\mathbb{Y}_6) > \hat{S}(\mathbb{Y}_4)$	$\mathbb{Y}_1 = \mathbb{Y}_2 = \mathbb{Y}_5 >$ $\mathbb{Y}_3 > \mathbb{Y}_7 > \mathbb{Y}_6 > \mathbb{Y}_4$

4 Two-person probabilistic multiplicative unbalanced linguistic cloud games (PM-ULCG)

In uncertain situations, the probabilistic unbalanced linguistic matrix game with payoffs as clouds is introduced in this section to model uncertain cloud payoff values. The payoff matrix $\widetilde{A}$ with probabilistic unbalanced linguistic information can be denoted as:

Player I	Player II			
	Strategy 1	Strategy 2	$\cdots$	Strategy s
Strategy 1	$\widetilde{a}_{11}$	$\widetilde{a}_{12}$	$\cdots$	$\widetilde{a}_{1s}$
Strategy 2	$\widetilde{a}_{21}$	$\widetilde{a}_{12}$	$\cdots$	$\widetilde{a}_{2s}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Strategy r	$\widetilde{a}_{r1}$	$\widetilde{a}_{r2}$	$\cdots$	$\widetilde{a}_{rs}$

where each $\widetilde{a}_{ij} \in LS(p)$ is a probabilistic unbalanced linguistic term.

4.1 PM-ULCG: definition and meaning

A two-person constant sum probabilistic multiplicative unbalanced linguistic cloud game is defined as $(S^r, S^s, \mathbb{LS}(p), \widetilde{A})$ where S^r and S^s are the strategy sets of Player I and Player II, respectively. $\mathbb{LS}(p)$ is the set of PM-UL terms having associated probabilities. Now to deal with the uncertainties like randomness and fuzziness underlying the PM-UL terms $\widetilde{a}_{ij}$, we first convert the payoffs in the form of PULC. Hence, finally, $\widetilde{A}$ is the payoff matrix of Player I corresponding to Player II given as

$$\widetilde{A} = \begin{pmatrix} \widetilde{a}_{11} & \widetilde{a}_{12} & \cdots & \widetilde{a}_{1s} \\ \widetilde{a}_{21} & \widetilde{a}_{22} & \cdots & \widetilde{a}_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{a}_{r1} & \widetilde{a}_{r2} & \cdots & \widetilde{a}_{rs} \end{pmatrix}$$

where each $\widetilde{a}_{ij} = (p_t EX_{a^t}, p_t EN_{a^t}, p_t HE_{a^t})$ is a PULC. On the contrary, $neg(\widetilde{A})$ is the payoff matrix of Player II, where $neg(\widetilde{A}) = neg(\widetilde{a}_{ij})_{r\times s}$, where $neg(\widetilde{a}_{ij})$ is as defined in Definition 10.

Conventionally, Player I is the row player and wants to maximize his payoff, while column player II wants to minimize row player's payoff so as to maximize his own payoff. The rows are pure strategies for Player I and columns for Player II.

$$v^- = \max_{i=1,\ldots,r} \min_{j=1,\ldots,s} \widetilde{a}_{ij}$$

called the lower value of the game, and is the minimum amount that Player I is guaranteed to receive. Here $\widetilde{a}_{ij}$ is a PM-UL term. On the other hand,

$$\omega^- = \max_{i=1,\ldots,r} \min_{j=1,\ldots,s} neg(\widetilde{a}_{ij})$$

is the minimum amount that Player *II* is guaranteed to receive. Obviously,

$$\upsilon^+ = neg(\omega^-) = \min_{j=1,\ldots,s} \max_{i=1,\ldots,r} \widetilde{a}_{ij}$$

is the upper value of the game and is the guaranteed maximum amount that Player *II* can lose. The game is said to have a value $\upsilon = (EX_{a^i}, EN_{a^i}, HE_{a^i})$ if $\upsilon^- = \upsilon^+ = \upsilon$, also called the saddle point in pure strategies. However, in most two-person zero sum games a saddle point in pure strategies will not exist because that would interpret that the players must always play in a similar manner. But that is not always possible as in such case, the opposing player will be able to figure out the player's strategy in advance.

Definition 14 A mixed cloud strategy is a vector $\varrho = (\varrho_1, \varrho_2, \ldots, \varrho_r)$ for Player *I* and $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_s)$ for Player *II*, where

$$\varrho_i \geq 0, \varrho_i = (EX_{a^i}, EN_{a^i}, HE_{a^i}) \text{ so that } \sum_{i=1}^{r} \varrho_i = 1 \implies \sum_{i=1}^{r} EX_{a^i} = 1,$$

$$\sqrt{\sum_{i=1}^{r} {EN_{a^i}}^2} = 1, \text{ and } \sqrt{\sum_{i=1}^{r} {HE_{a^i}}^2} = 1$$

and,

$$\varphi_j \geq 0, \varphi_j = (EX_{a^j}, EN_{a^j}, HE_{a^j}) \text{ so that } \sum_{j=1}^{s} \varphi_j = 1 \implies \sum_{j=1}^{s} EX_{a^j} = 1,$$

$$\sqrt{\sum_{j=1}^{s} {EN_{a^j}}^2} = 1, \text{ and } \sqrt{\sum_{j=1}^{s} {HE_{a^j}}^2} = 1$$

The set of mixed strategies having k components is denoted as

$$S^k = \left\{ (z_1, z_2, \ldots, z_k) | z_i \geq 0, \sum_{i=1}^{k} EX_{a^i} = 1, \sqrt{\sum_{i=1}^{k} {EN_{a^i}}^2} = 1, \text{ and } \sqrt{\sum_{i=1}^{k} {HE_{a^i}}^2} = 1 \right\}$$

such that $\varrho \in S^r$ and $\varphi \in S^s$.

Since it is mentioned that a PM-ULCG has a saddle point in pure strategies if $\upsilon^- = \upsilon^+$. However, in practical real-life scenarios, this is not always the case that saddle point exists. Hence in that case, the following inequality holds in general.

Theorem 1 *Suppose v^- and v^+ be the lower and upper values of a PM-ULCG, respectively, and let both the values exist. Then $v^- \leq v^+$.*

Proof For any column j, we know that for any fixed row i, by Definition 13

$$\min_{j=1,\ldots,s} \widetilde{a}_{ij} \leq \widetilde{a}_{ij}$$

Taking max over $i = 1, \ldots, r$ both the sides,

$$v^- = \max_{i=1,\ldots,r} \min_{j=1,\ldots,s} \widetilde{a}_{ij} \leq \max_{i=1,\ldots,r} \widetilde{a}_{ij}$$

This is true for any column $j = 1, \ldots, s$. The LHS is just a number independent of i and j and is smaller than RHS for any j. This means that $v^- \leq \min_{j=1,\ldots,s} \max_{i=1,\ldots,r} \widetilde{a}_{ij} = v^+$. This proves the theorem. □

In crisp games, the expected payoff of the player is presented in the form of mathematical expectation; but the uncertainty pertaining in the PM-ULCG does not allow the expected payoffs to be defined simply as in crisp games. Thus, it is necessary to define PM-UL cloud expected payoffs.

Definition 15 Let $\varrho \in S^r$ and $\varphi \in S^s$ be the mixed strategies associated with Player *I* and Player *II*, respectively. Then, the PM-UL cloud expected payoff of Player *I* is defined as:

$$E(\varrho, \varphi) = \bigoplus_{i=1}^{r} \left(\varrho_i \left(\bigoplus_{j=1}^{s} \widetilde{a}_{ij} \varphi_j \right) \right)$$

where $\widetilde{a}_{ij} = (p_t EX_{a^t}, p_t EN_{a^t}, p_t HE_{a^t})$, $t = -n, \ldots, m$ is a PULC.

$$\implies E(\varrho, \varphi) = \left(\sum_{i=1}^{r} \sum_{j=1}^{s} \varrho_i \varphi_j p_t EX_{a^t}, \sqrt{\sum_{i=1}^{r} \sum_{j=1}^{s} \varrho_i \varphi_j \left(p_t EN_{a^t} \right)^2}, \sqrt{\sum_{i=1}^{r} \sum_{j=1}^{s} \varrho_i \varphi_j \left(p_t HE_{a^t} \right)^2} \right) \tag{10}$$

and this expected payoff is further a PULC. Similarly, the PM-UL cloud expected payoff of Player *II* is given by $Neg(E(\varrho, \varphi))$.

4.2 PM-ULCG: methodology and solution

This section discusses the probabilistic unbalanced linguistic cloud linear programming (PULCLP) method for solving matrix games problems having payoffs in the form of PM-UL terms.

Suppose $\mathbb{LS}(p) = \{t_{a^i}(p_i) | i = -n, \ldots, m\}$ where n and m are positive integers and $a > 1$ be the linguistic set having PM-UL terms with probabilities

associated with them; $\widetilde{A} = [\widetilde{a}_{rs}]$ be the PULC matrix so that each $\widetilde{a}_{ij}$ represents a PULC written as $\widetilde{a}_{ij} = (p_t EX_{a^t}, p_t EN_{a^t}, p_t HE_{a^t})$. Let υ^- refers to the lower value of the game corresponding to Player I and let $\varrho \in S^r$ be the mixed strategies of Player I. Now, the PM-UL cloud expected payoff of Player I for s^{th} strategy of Player II is the weighted average of $\widetilde{a}_{ij}$, i.e. $\bigoplus_{j=1}^{s} \left(\varrho_i \widetilde{a}_{ij}\right)$ where $\widetilde{a}_{ij} = (p_t EX_{a^t}, p_t EN_{a^t}, p_t HE_{a^t})$, $t = -n, \ldots, m$ is a PULC such that $i = 1, 2, \ldots, r$. Analogously, υ^+ is the maximum expected payoff of Player II, $\varphi \in S^s$ is its mixed strategy, and its expected payoff against Player I's r^{th} strategy is given by $\bigoplus_{i=1}^{r} \left(\varphi_i \widetilde{a}_{ij}\right)$ where $j = 1, 2, \ldots, s$.

For finding the expected payoffs and values of both the players in terms of PULCs, we first convert each PM-UL payoff to a PULC.

Hence, the PULCLP for Player I can be given as

$$\max \upsilon^- \tag{P1}$$

$$\text{s.t.} \begin{cases} \bigoplus_{i=1}^{r} \widetilde{a}_{ij} \varrho_i \geq \upsilon^- \; j = 1, 2, \ldots, s \\ \text{where} \\ \varrho_i \geq 0, \; \sum_{i=1}^{r} \varrho_i = 1. \end{cases}$$

and for Player II, the problem is

$$\max \omega^- \tag{P2}$$

$$\text{s.t.} \begin{cases} \bigoplus_{j=1}^{s} neg(\widetilde{a}_{ij}) \varphi_j \geq \omega^- \; i = 1, 2, \ldots, r \\ \text{where} \\ \varphi_j \geq 0, \; \sum_{j=1}^{s} \varphi_j = 1. \end{cases}$$

In above problems, each $\widetilde{a}_{ij}$ is a PULC. Further, simplifying P2 and using the properties of negation operator, the problem for Player II becomes

$$\begin{array}{l} \min \quad \upsilon^+ \\ \text{s.t.} \begin{cases} \bigoplus_{j=1}^{s} \widetilde{a}_{ij} \varphi_j \leq \upsilon^+ \; i = 1, 2, \ldots, r \\ \text{where} \\ \varphi_j \geq 0, \; \sum_{j=1}^{s} \varphi_j = 1. \end{cases} \end{array} \tag{P3}$$

A unified mechanism for solving a probabilistic unbalanced linguistic game for Player I is given below in the form of a computation procedure.

Step 1: First normalize the payoff matrix for the row player (i.e. normalize the payoff entries row-wise) so that the sum of probabilities in each row is 1.

Step 2: Convert each payoff to a PULC using the procedure given in Sect. 3.1 to deal with uncertainty in a better way.

Step 3: Since each payoff is a PULC now, suppose the expected value of Player I be $\upsilon^- = (v_{EX}{}^-, v_{EN}{}^-, v_{HE}{}^-)$ such that the PULCLP becomes

$$\max \quad \upsilon^{-} \implies (\max \upsilon_{EX}{}^{-}, \min \upsilon_{EN}{}^{-}, \min \upsilon_{HE}{}^{-})$$
$$\text{s.t.} \begin{cases} \bigoplus_{i=1}^{r} \widetilde{a}_{ij} \varrho_i \geq \upsilon^{-} \; j = 1, 2, \ldots, s \\ \text{where} \\ \varrho_i \geq 0, \sum_{i=1}^{r} \varrho_i = 1. \end{cases} \tag{P4}$$

Here each $\widetilde{a}_{ij} = (p_t EX_{a^t}, p_t EN_{a^t}, p_t HE_{a^t})$.

Step 4: From here, the PM-ULCG splits into three individual games, called the expectation game, entropy game, and hyperentropy game given as: Expectation Game

$$\max \quad \upsilon_{EX}{}^{-}$$
$$\text{s.t.} \begin{cases} \sum_{i=1}^{r} (p_t EX_{a^t})_{ij} \varrho_i \geq \upsilon_{EX}{}^{-} \; j = 1, 2, \ldots, s \\ \text{where} \\ \varrho_i \geq 0, \sum_{i=1}^{r} \varrho_i = 1. \end{cases} \tag{EX:1}$$

Entropy Game

$$\min \quad \upsilon_{EN}{}^{-}$$
$$\text{s.t.} \begin{cases} \sqrt{\sum_{i=1}^{r} [(p_t EN_{a^t})_{ij}]^2} \varrho_i \leq \upsilon_{EN}{}^{-} \; j = 1, 2, \ldots, s \\ \text{where} \\ \varrho_i \geq 0, \sum_{i=1}^{r} \varrho_i = 1. \end{cases} \tag{EN:1}$$

Hyperentropy Game

$$\min \quad \upsilon_{HE}{}^{-}$$
$$\text{s.t.} \begin{cases} \sqrt{\sum_{i=1}^{r} [(p_t HE_{a^t})_{ij}]^2} \varrho_i \leq \upsilon_{HE}{}^{-} \; j = 1, 2, \ldots, s \\ \text{where} \\ \varrho_i \geq 0, \sum_{i=1}^{r} \varrho_i = 1. \end{cases} \tag{HE:1}$$

Step 5: Proceeding further, Expectation Game Setting $X_i = \frac{\varrho_i}{\upsilon_{EX}{}^{-}}$ and letting $V_{EX} = \frac{1}{\upsilon_{EX}{}^{-}}$, the transformed linear programming problem becomes

$$\min \quad V_{EX} = \sum_{i=1}^{r} X_i$$
$$\text{s.t.} \begin{cases} \sum_{i=1}^{r} (p_t EX_{a^t})_{ij} X_i \geq 1 \; j = 1, 2, \ldots, s \\ \text{where } X_i \geq 0 \end{cases} \tag{EX Game PI}$$

Entropy Game Squaring both sides of the constraints and letting $(\upsilon_{EN}{}^{-})^2 = \upsilon_{EN}{}'$. Further setting $Y_i = \frac{\varrho_i}{\upsilon_{EN}{}'}$, and letting $V_{EN}{}' = \frac{1}{\upsilon_{EN}{}'} \upsilon_{EN}{}^{-} = \sqrt{\upsilon_{EN}{}'} = \frac{1}{\sqrt{V_{EN}{}'}}$
$\therefore \min \upsilon_{EN}{}^{-} \equiv \max \sqrt{V_{EN}{}'}$ $V_{EN}{}'$ being a positive quantity, we maximize $V_{EN}{}'$ instead of $\sqrt{V_{EN}{}'}$ and square the value later on. Thus, the final entropy linear programming problem becomes

$$\begin{aligned} &\max \quad {V_{EN}}' = \sum_{i=1}^{r} Y_i \\ &\text{s.t.} \begin{cases} \sum_{i=1}^{r} \left[(p_t EN_{a^t})_{ij}\right]^2 Y_i \le 1,\ j = 1, 2, \ldots, s \\ \text{where } Y_i \ge 0 \end{cases} \end{aligned} \tag{EN Game PI}$$

Hyperentropy Game Similar to entropy game, squaring both sides and letting $(\upsilon_{HE})^2 = {\upsilon_{HE}}'$. Further setting $Z_i = \frac{\varrho_i}{{\upsilon_{HE}}'}$, and suppose ${V_{HE}}' = \frac{1}{{\upsilon_{HE}}'} {\upsilon_{HE}}^- = \sqrt{{\upsilon_{HE}}'} = \frac{1}{\sqrt{{V_{HE}}'}} \therefore \min {\upsilon_{HE}}^- \equiv \max \sqrt{{V_{HE}}'}\ {V_{HE}}'$ being a positive quantity, we maximize ${V_{HE}}'$ instead of $\sqrt{{V_{HE}}'}$ and square the value later on. Thus, the final hyperentropy linear programming problem becomes

$$\begin{aligned} &\max \quad {V_{HE}}' \\ &\text{s.t.} \begin{cases} \sum_{i=1}^{r} \left[(p_t HE_{a^t})_{ij}\right]^2 Z_i \le 1,\ j = 1, 2, \ldots, s \\ \text{where } Z_i \ge 0 \end{cases} \end{aligned} \tag{HE Game PI}$$

Step 6: After the formulation of final LPPs each for expectation, entropy and hyperentropy, we solve them to get the PULC strategies and PULC value of each game. Combining all three, we get Lower value of the game: $\upsilon^- = ({\upsilon_{EX}}^-, {\upsilon_{EN}}^-, {\upsilon_{HE}}^-)$ and mixed strategies ϱ_i for each game such that $\sum_{i=1}^{r} \varrho_{i_{EX}} = 1,\ \sqrt{\sum_{i=1}^{r} \varrho_{i_{EN}}^2} = 1,\ \sqrt{\sum_{i=1}^{r} \varrho_{i_{HE}}^2} = 1$

Step 7: Substituting back to the original variables, we get the final results. Also, the final strategy to be played by the player may be decided by using score of each strategy. The higher the score, the better (optimum) is the strategy.

Figure 2 demonstrates the process of evaluation of a PM-ULCG problem for the row player.

The exact same procedure is to be followed for Player *II*, where the payoffs are $neg(\widetilde{a}_{ij})$. The expected value of Player *II* would be $\upsilon^+ = ({\upsilon_{EX}}^+, {\upsilon_{EN}}^+, {\upsilon_{HE}}^+)$ such that as per P3, the PULCLP becomes

$$\begin{aligned} &\min \quad \upsilon^+ \implies (\min {\upsilon_{EX}}^+, \max {\upsilon_{EN}}^+, \max {\upsilon_{HE}}^+) \\ &\text{s.t.} \begin{cases} \bigoplus_{j=1}^{s} \widetilde{a}_{ij} \varphi_j \le \upsilon^+ \ i = 1, 2, \ldots, r \\ \text{where} \\ \varrho_j \ge 0,\ \sum_{j=1}^{s} \varphi_j = 1. \end{cases} \end{aligned} \tag{P5}$$

and thus, the PULCLP splits into three problems given as: Expectation Game

$$\begin{aligned} &\min \quad {\upsilon_{EX}}^+ \\ &\text{s.t.} \begin{cases} \sum_{j=1}^{s} (p_t EX_{a^t})_{ij} \varrho_j \le {\upsilon_{EX}}^+ \ i = 1, 2, \ldots, r \\ \text{where} \\ \varrho_j \ge 0,\ \sum_{j=1}^{s} \varphi_j = 1. \end{cases} \end{aligned} \tag{EX Game PII}$$

Entropy Game

$$\max \quad {v_{EN}}^{+}$$
$$\text{s.t.} \begin{cases} \sqrt{\sum_{j=1}^{s} \left[(p_t EN_{a^t})_{ij}\right]^2 \varrho_j} \geq {v_{EN}}^{+} \; i = 1, 2, \ldots, r \\ \text{where} \\ \varrho_j \geq 0, \; \sum_{j=1}^{s} \varphi_j = 1. \end{cases} \qquad \text{(EN Game PII)}$$

Hyperentropy Game

$$\max \quad {v_{HE}}^{+}$$
$$\text{s.t.} \begin{cases} \sqrt{\sum_{j=1}^{s} \left[(p_t HE_{a^t})_{ij}\right]^2 \varrho_j} \geq {v_{HE}}^{+} \; i = 1, 2, \ldots, r \\ \text{where} \\ \varrho_j \geq 0, \; \sum_{j=1}^{s} \varphi_j = 1. \end{cases} \qquad \text{(HE Game PII)}$$

Then, solving just as in Step 5 mentioned above, we combine the strategies and values obtained from all three games so that we get upper value of the game, i.e. $v^{+} = ({v_{EX}}^{+}, {v_{EN}}^{+}, {v_{HE}}^{+})$.

We illustrate the defined methodology using an example.

4.3 Numerical illustration

Example 3 An investor wishes to invest his money in the stocks of some IT companies. He does so by analysing the average stock closing values for the past 6 years 2017–2022. The possible alternatives under consideration are the five IT companies: Infosys, Wipro, TCS, HCL Technologies, and Larsen & Toubro Infotech. The following linguistic variables represent the stock values:

$$LS = \{t_{2^{-2}} : \text{Very Low (VL)}, t_{2^{-1}} : \text{Low (L)}, t_{2^{0}} : \text{Average (Avg)}, t_{2^{1}} : \text{High (H)}, t_{2^{2}} : \text{Very High (VH)}\}$$

Table 2 provides the values of stocks of the IT companies for the financial years 2017-2022. These data are taken from https://www.moneycontrol.commoneycontrol.

Based on the values listed in Table 2, the variations that occurred in the stock values of different IT companies are evaluated in terms of PM-ULTS (just an interpretation). The interpretation here is done by supposing that the values less than 300 assume the linguistic term “VL”, 300–900 assume “L”, 900–1500 as “Avg”, 1500–2100 as “H” and greater than 2100 assume the linguistic term “VH”. Also, the probabilistic values are interpreted based on the deviation of numeric values from the original linguistic terms. Thus, the probabilistic unbalanced linguistic stock data for the years 2017–2022 for the above-mentioned IT companies are provided in the following “IT company versus year" matrix:

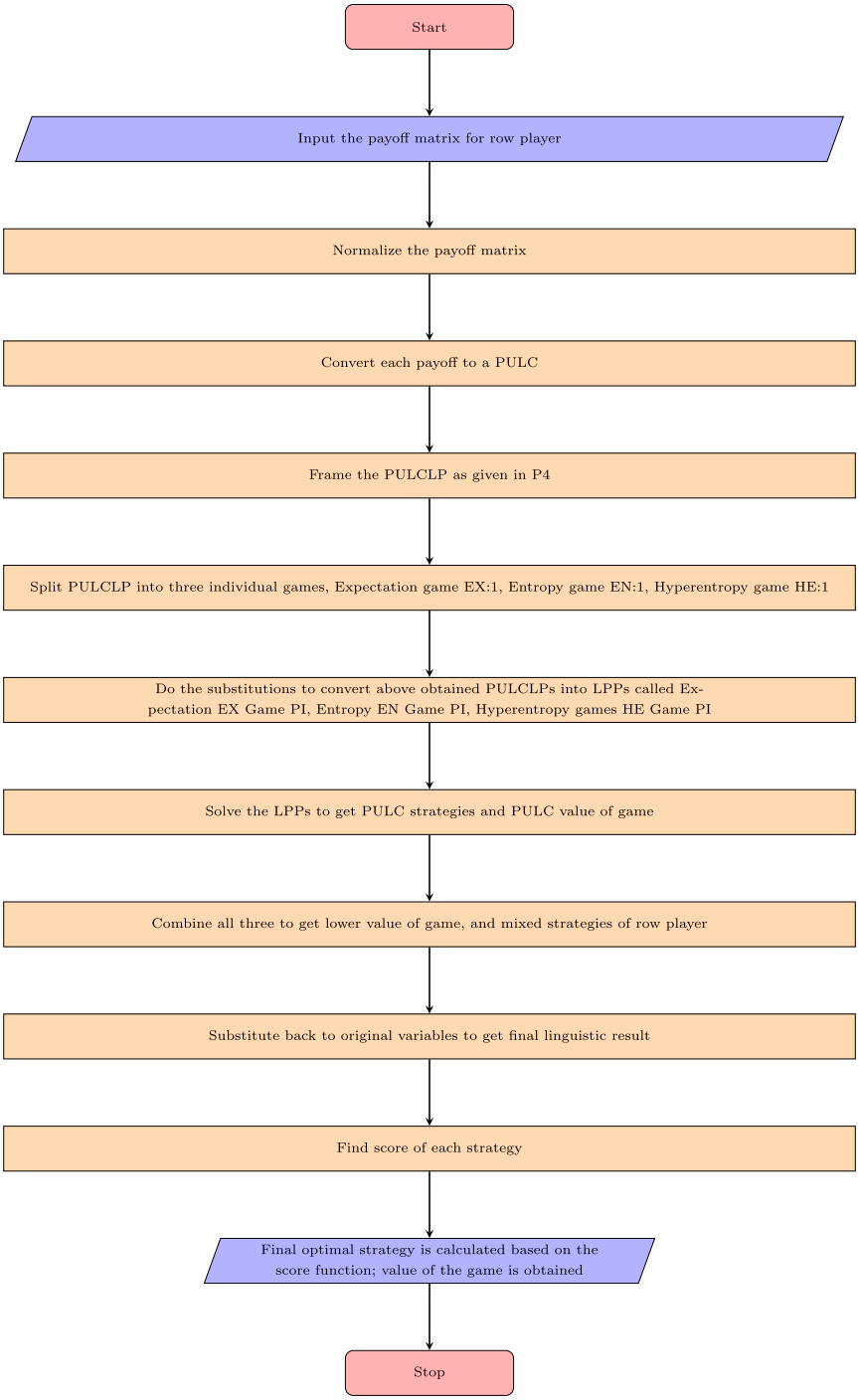


Fig. 2 Flowchart for evaluation of PM-ULCG problem for row player

$$\widetilde{A} = \begin{array}{c} \\ \text{Infosys} \\ \text{Wipro} \\ \text{TCS} \\ \text{HCL Technologies} \\ \text{Larsen \& Toubro Infotech} \end{array} \begin{array}{c} \begin{array}{cccccc} 2017 & 2018 & 2019 & 2020 & 2021 & 2022 \end{array} \\ \begin{bmatrix} t_{2^{-1}}(0.4) & t_{2^{-1}}(0.6) & t_{2^{-1}}(0.7) & t_{2^0}(0.6) & t_{2^1}(0.7) & t_{2^1}(0.1) \\ t_{2^{-2}}(0.2) & t_{2^{-2}}(0.2) & t_{2^{-2}}(0.2) & t_{2^{-1}}(0.2) & t_{2^{-1}}(0.7) & t_{2^{-1}}(0.2) \\ t_{2^0}(0.7) & t_{2^1}(0.7) & t_{2^2}(0.1) & t_{2^2}(0.2) & t_{2^2}(0.4) & t_{2^2}(0.4) \\ t_{2^{-1}}(0.3) & t_{2^{-1}}(0.4) & t_{2^{-1}}(0.5) & t_{2^0}(0.1) & t_{2^0}(0.7) & t_{2^0}(0.4) \\ t_{2^0}(0.4) & t_{2^1}(0.4) & t_{2^1}(0.5) & t_{2^2}(0.4) & t_{2^2}(1) & t_{2^1}(0.8) \end{bmatrix} \end{array}$$

The payoff $\widetilde{a}_{ij}$ of Player *I* represents the response to strategy *i* of Player *I* for the case when Player *II* selects strategy *j* in order to win the game. Also, a probabilistic linguistic payoff signifies that any strategy played by a player can be evaluated on an unbalanced linguistic scale.

Next, we apply our methodology to deal with PM-UL information to analyse which company has had a higher growth in share market from financial year 2017 to 2022. On the basis of this analysis, one can interpret which company to invest in, for future purpose. Now first we normalize the payoff matrix. So, the normalized payoff matrix may be given as:

$$\widetilde{A} = \begin{array}{c} \\ \text{Infosys} \\ \text{Wipro} \\ \text{TCS} \\ \text{HCL Technologies} \\ \text{Larsen \& Toubro Infotech} \end{array} \begin{array}{c} \begin{array}{cccccc} 2017 & 2018 & 2019 & 2020 & 2021 & 2022 \end{array} \\ \begin{bmatrix} t_{2^{-1}}(0.13) & t_{2^{-1}}(0.194) & t_{2^{-1}}(0.225) & t_{2^0}(0.194) & t_{2^1}(0.225) & t_{2^1}(0.032) \\ t_{2^{-2}}(0.118) & t_{2^{-2}}(0.118) & t_{2^{-2}}(0.118) & t_{2^{-1}}(0.118) & t_{2^{-1}}(0.41) & t_{2^{-1}}(0.118) \\ t_{2^0}(0.28) & t_{2^1}(0.28) & t_{2^2}(0.04) & t_{2^2}(0.08) & t_{2^2}(0.16) & t_{2^2}(0.16) \\ t_{2^{-1}}(0.125) & t_{2^{-1}}(0.167) & t_{2^{-1}}(0.208) & t_{2^0}(0.041) & t_{2^0}(0.292) & t_{2^0}(0.167) \\ t_{2^0}(0.114) & t_{2^1}(0.114) & t_{2^1}(0.144) & t_{2^2}(0.114) & t_{2^2}(0.286) & t_{2^1}(0.228) \end{bmatrix} \end{array}$$

The ambiguity and volatility of stock values are represented by the dispersion of cloud droplets. So we convert the probabilistic unbalanced linguistic terms to PULCs so as to choose a specific alternative. Hence, we use the procedure given in Sect. 3.1 to convert the payoffs to PULCs, and thus, the PULC payoff matrix is given as:

$$\widetilde{A}_{PULC} = \begin{array}{c} \\ \text{Infosys} \\ \text{Wipro} \\ \text{TCS} \\ \text{HCL Technologies} \\ \text{Larsen \& Toubro Infotech} \end{array} \begin{array}{c} \begin{array}{cccccc} 2017 & 2018 & 2019 & 2020 & 2021 & 2022 \end{array} \\ \begin{bmatrix} (0.6298, 1.76033, 0.6987) & (0.9399, 2.6269, 1.0427) & (1.0901, 3.0467, 1.2093) & (1.6112, 2.2799, 1.42706) & (2.6473, 3.0467, 1.2093) & (0.3765, 0.4333, 0.172) \\ (0, 1.854, 0.3506) & (0, 1.854, 0.3506) & (0, 1.854, 0.3506) & (0.5717, 1.5978, 0.6342) & (1.9864, 5.5518, 2.203) & (0.5717, 1.5978, 0.6342) \\ (2.3255, 3.2907, 2.0596) & (3.2944, 3.7914, 1.505) & (0.6284, 0.6284, 0.1188) & (1.2569, 1.2569, 0.2376) & (2.5139, 2.5139, 0.4753) & (2.5139, 2.5139, 0.4753) \\ (0.6056, 1.6926, 0.6718) & (0.8091, 2.2613, 0.8976) & (1.0077, 2.8165, 1.118) & (0.3405, 0.4818, 0.3015) & (2.4252, 3.4317, 2.1479) & (1.387, 1.9626, 1.2284) \\ (0.9468, 1.3397, 0.8385) & (1.3413, 1.5436, 0.6127) & (1.6943, 1.9499, 0.774) & (1.7911, 1.7911, 0.3387) & (4.4936, 4.4936, 0.8497) & (2.6826, 3.0873, 1.2255) \end{bmatrix} \end{array}$$

Solving this game problem for Player *I*, i.e. for “IT companies", we get three separate games as expectation game, entropy game and hyperentropy game given as:

Expectation Game

Table 2 Closing stock prices of IT companies for different financial years

	2017	2018	2019	2020	2021	2022
Infosys	521.03	658.95	731.15	1255.8	1887.75	1508.2
Wipro	235.69	248.14	245.8	386.25	715.35	392.75
TCS	1350.6	1893.05	2161.7	2862.75	3738.35	3256.7
HCL Technologies	445.25	482.18	568.1	946.15	1319.1	1039.3
Larsen & Toubro Infotech	1118.85	1731.2	1750.5	3659.5	7332	2085.8

$$\max \quad {v_{EX}}^{-}$$
$$\text{s.t.} \begin{cases} 0.6298\varrho_1 + 0\varrho_2 + 2.3255\varrho_3 + 0.6056\varrho_4 + 0.9468\varrho_5 \geq {v_{EX}}^{-} \\ 0.9399\varrho_1 + 0\varrho_2 + 3.29448\varrho_3 + 0.8091\varrho_4 + 1.3413\varrho_5 \geq {v_{EX}}^{-} \\ 1.0901\varrho_1 + 0\varrho_2 + 0.6284\varrho_3 + 1.0077\varrho_4 + 1.6943\varrho_5 \geq {v_{EX}}^{-} \\ 1.6112\varrho_1 + 0.5717\varrho_2 + 1.2569\varrho_3 + 0.3405\varrho_4 + 1.7911\varrho_5 \geq {v_{EX}}^{-} \\ 2.6473\varrho_1 + 1.9864\varrho_2 + 2.5139\varrho_3 + 2.4252\varrho_4 + 4.4936\varrho_5 \geq {v_{EX}}^{-} \\ 0.3765\varrho_1 + 0.5717\varrho_2 + 2.5139\varrho_3 + 1.387\varrho_4 + 2.6826\varrho_5 \geq {v_{EX}}^{-} \\ \text{where } \varrho_i \geq 0,\ \sum_{i=1}^{5} \varrho_i = 1 \end{cases}$$

Setting $X_i = \frac{\varrho_i}{{v_{EX}}^{-}}$, the problem becomes

$$\min \quad V_{EX} = X_1 + X_2 + X_3 + X_4 + X_5$$
$$\text{s.t.} \begin{cases} 0.6298X_1 + 0X_2 + 2.3255X_3 + 0.6056X_4 + 0.9468X_5 \geq 1 \\ 0.9399X_1 + 0X_2 + 3.29448X_3 + 0.8091X_4 + 1.3413X_5 \geq 1 \\ 1.0901X_1 + 0X_2 + 0.6284X_3 + 1.0077X_4 + 1.6943X_5 \geq 1 \\ 1.6112X_1 + 0.5717X_2 + 1.2569X_3 + 0.3405X_4 + 1.7911X_5 \geq 1 \\ 2.6473X_1 + 1.9864X_2 + 2.5139X_3 + 2.4252X_4 + 4.4936X_5 \geq 1 \\ 0.3765X_1 + 0.5717X_2 + 2.5139X_3 + 1.387X_4 + 2.6826X_5 \geq 1 \\ \text{where } X_i \geq 0,\ i = 1, \ldots, 5 \end{cases}$$

Solving the above LPP, we get the solution as: $X_1 = 0$, $X_2 = 0$, $X_3 = 0.223$, $X_4 = 0$, $X_5 = 0.508$, $V_{EX} = 0.731$. Substituting back to the original variables, we get the optimal solution of the expectation game problem as: ${v_{EX}}^{-} = \frac{1}{V_{EX}} = 1.36$, $\varrho_1 = 0$, $\varrho_2 = 0$, $\varrho_3 = 0.305$, $\varrho_4 = 0$, $\varrho_5 = 0.695$. Hence, the strategies are (0, 0, 0.305, 0, 0.695) and expectation value is 1.36. Entropy Game

$$\min \quad {v_{EN}}^{-}$$
$$\text{s.t.} \begin{cases} \sqrt{(1.7603)^2\varrho_1 + (1.854)^2\varrho_2 + (3.2907)^2\varrho_3 + (1.6926)^2\varrho_4 + (1.3397)^2\varrho_5} \leq {v_{EN}}^{-} \\ \sqrt{(2.6269)^2\varrho_1 + (1.854)^2\varrho_2 + (3.7914)^2\varrho_3 + (2.2613)^2\varrho_4 + (1.5436)^2\varrho_5} \leq {v_{EN}}^{-} \\ \sqrt{(3.0467)^2\varrho_1 + (1.854)^2\varrho_2 + (0.6285)^2\varrho_3 + (2.8165)^2\varrho_4 + (1.95)^2\varrho_5} \leq {v_{EN}}^{-} \\ \sqrt{(2.2799)^2\varrho_1 + (1.5978)^2\varrho_2 + (1.2569)^2\varrho_3 + (0.482)^2\varrho_4 + (1.791)^2\varrho_5} \leq {v_{EN}}^{-} \\ \sqrt{(3.0467)^2\varrho_1 + (5.5518)^2\varrho_2 + (2.5139)^2\varrho_3 + (3.4317)^2\varrho_4 + (4.4936)^2\varrho_5} \leq {v_{EN}}^{-} \\ \sqrt{(0.4333)^2\varrho_1 + (1.5978)^2\varrho_2 + (2.5139)^2\varrho_3 + (1.9626)^2\varrho_4 + (3.0873)^2\varrho_5} \leq {v_{EN}}^{-} \\ \text{where } \varrho_i \geq 0,\ i = 1, \ldots, 5\ \sum_{i=1}^{5} \varrho_i = 1 \end{cases}$$

Squaring both sides, and letting $({v_{EN}}^{-})^2 = {v_{EN}}'$, we get

$$\min \quad \sqrt{{v_{EN}}'} \equiv \min {v_{EN}}'$$

$$\text{s.t.} \begin{cases} 3.0987\varrho_1 + 3.4374\varrho_2 + 10.8287\varrho_3 + 2.8649\varrho_4 + 1.7948\varrho_5 \le {v_{EN}}' \\ 6.9\varrho_1 + 3.437\varrho_2 + 14.375\varrho_3 + 5.113\varrho_4 + 2.382\varrho_5 \le {v_{EN}}' \\ 9.282\varrho_1 + 3.437\varrho_2 + 0.395\varrho_3 + 7.933\varrho_4 + 3.802\varrho_5 \le {v_{EN}}' \\ 5.198\varrho_1 + 2.553\varrho_2 + 1.579\varrho_3 + 0.232\varrho_4 + 3.208\varrho_5 \le {v_{EN}}' \\ 9.282\varrho_1 + 30.822\varrho_2 + 6.319\varrho_3 + 11.653\varrho_4 + 20.192\varrho_5 \le {v_{EN}}' \\ 0.187\varrho_1 + 2.553\varrho_2 + 6.319\varrho_3 + 3.851\varrho_4 + 9.531\varrho_5 \le {v_{EN}}' \\ \text{where } \varrho_i \ge 0,\ i = 1, \ldots, 5 \sum_{i=1}^{5} \varrho_i = 1 \end{cases}$$

Setting $Y_i = \frac{\varrho_i}{{v_{EN}}'}$ and ${V_{EN}}' = \frac{1}{{v_{EN}}'}$, hence the final entropy LPP becomes

$$\max \quad {V_{EN}}' = Y_1 + Y_2 + Y_3 + Y_4 + Y_5$$

$$\text{s.t.} \begin{cases} 3.0987Y_1 + 3.4374Y_2 + 10.8287Y_3 + 2.8649Y_4 + 1.7948Y_5 \le 1 \\ 6.9Y_1 + 3.437Y_2 + 14.375Y_3 + 5.113Y_4 + 2.382Y_5 \le 1 \\ 9.282Y_1 + 3.437Y_2 + 0.395Y_3 + 7.933Y_4 + 3.802Y_5 \le 1 \\ 5.198Y_1 + 2.553Y_2 + 1.579Y_3 + 0.232Y_4 + 3.208Y_5 \le 1 \\ 9.282Y_1 + 30.822Y_2 + 6.319Y_3 + 11.653Y_4 + 20.192Y_5 \le 1 \\ 0.187Y_1 + 2.553Y_2 + 6.319Y_3 + 3.851Y_4 + 9.531Y_5 \le 1 \\ \text{where } Y_i \ge 0,\ i = 1, \ldots, 5 \end{cases}$$

Solving the above LPP, the solution comes out to be $Y_1 = 0.08968$, $Y_2 = 0$, $Y_3 = 0.0265$, $Y_4 = 0$, $Y_5 = 0$. Further, ${V_{EN}}' = 0.1162$. $\therefore {v_{EN}}^- = 2.93$ and the strategies are (0.7717, 0, 0.2282, 0, 0). On a similar note, Hyperentropy Game

$$\min \quad {v_{HE}}^-$$

$$\text{s.t.} \begin{cases} \sqrt{(0.6987)^2\varrho_1 + (0.3506)^2\varrho_2 + (2.0596)^2\varrho_3 + (0.6718)^2\varrho_4 + (0.8385)^2\varrho_5} \le {v_{HE}}^- \\ \sqrt{(1.0427)^2\varrho_1 + (0.3506)^2\varrho_2 + (1.505)^2\varrho_3 + (0.8976)^2\varrho_4 + (0.6127)^2\varrho_5} \le {v_{HE}}^- \\ \sqrt{(1.2093)^2\varrho_1 + (0.3506)^2\varrho_2 + (0.1188)^2\varrho_3 + (1.118)^2\varrho_4 + (0.774)^2\varrho_5} \le {v_{HE}}^- \\ \sqrt{(1.427)^2\varrho_1 + (0.6342)^2\varrho_2 + (0.2376)^2\varrho_3 + (0.3015)^2\varrho_4 + (0.3387)^2\varrho_5} \le {v_{HE}}^- \\ \sqrt{(1.2093)^2\varrho_1 + (2.203)^2\varrho_2 + (0.4754)^2\varrho_3 + (2.1479)^2\varrho_4 + (0.8479)^2\varrho_5} \le {v_{HE}}^- \\ \sqrt{(0.172)^2\varrho_1 + (0.6342)^2\varrho_2 + (0.4754)^2\varrho_3 + (1.2284)^2\varrho_4 + (1.2255)^2\varrho_5} \le {v_{HE}}^- \\ \text{where } \varrho_i \ge 0,\ i = 1, \ldots, 5 \sum_{i=1}^{5} \varrho_i = 1 \end{cases}$$

Squaring both sides and letting $({v_{HE}}^-)^2 = {v_{HE}}'$

$$\min \quad \sqrt{{\upsilon_{HE}}'} \equiv \min {\upsilon_{HE}}'$$
$$\text{s.t.} \begin{cases} 0.4882\varrho_1 + 0.1229\varrho_2 + 4.2423\varrho_3 + 0.4513\varrho_4 + 0.7031\varrho_5 \le {\upsilon_{HE}}' \\ 1.0872\varrho_1 + 0.1229\varrho_2 + 2.265\varrho_3 + 0.8056\varrho_4 + 0.3755\varrho_5 \le {\upsilon_{HE}}' \\ 1.4624\varrho_1 + 0.1229\varrho_2 + 0.0141\varrho_3 + 1.2499\varrho_4 + 0.599\varrho_5 \le {\upsilon_{HE}}' \\ 2.036\varrho_1 + 0.4022\varrho_2 + 0.0564\varrho_3 + 0.0909\varrho_4 + 0.1147\varrho_5 \le {\upsilon_{HE}}' \\ 1.4624\varrho_1 + 4.8532\varrho_2 + 0.226\varrho_3 + 4.6134\varrho_4 + 0.7219\varrho_5 \le {\upsilon_{HE}}' \\ 0.029\varrho_1 + 0.4022\varrho_2 + 0.226\varrho_3 + 1.5089\varrho_4 + 1.5018\varrho_5 \le {\upsilon_{HE}}' \\ \text{where } \varrho_i \ge 0,\ i = 1, \ldots, 5\ \sum_{i=1}^{5} \varrho_i = 1 \end{cases}$$

Setting $Z_i = \frac{\varrho_i}{{\upsilon_{HE}}'}$ and ${V_{HE}}' = \frac{1}{{\upsilon_{HE}}'}$, the final hyperentropy LPP becomes

$$\max \quad {V_{HE}}'$$
$$\text{s.t.} \begin{cases} 0.4882Z_1 + 0.1229Z_2 + 4.2423Z_3 + 0.4513Z_4 + 0.7031Z_5 \le 1 \\ 1.0872Z_1 + 0.1229Z_2 + 2.265Z_3 + 0.8056Z_4 + 0.3755Z_5 \le 1 \\ 1.4624Z_1 + 0.1229Z_2 + 0.0141Z_3 + 1.2499Z_4 + 0.599Z_5 \le 1 \\ 2.036Z_1 + 0.4022Z_2 + 0.0564Z_3 + 0.0909Z_4 + 0.1147Z_5 \le 1 \\ 1.4624Z_1 + 4.8532Z_2 + 0.226Z_3 + 4.6134Z_4 + 0.7219Z_5 \le 1 \\ 0.029Z_1 + 0.4022Z_2 + 0.226Z_3 + 1.5089Z_4 + 1.5018Z_5 \le 1 \\ \text{where } Z_i \ge 0,\ i = 1, \ldots, 5 \end{cases}$$

Solving the above LPP, solution come out as $Z_1 = 0.3514$, $Z_2 = 0, Z_3 = 0.0883$, $Z_4 = 0$, $Z_5 = 0.6458$, ${V_{HE}}' = 1.0854$. Substituting back to original variables, the hyperentropy value ${\upsilon_{HE}}^- = 0.96$, and the strategies are (0.3237, 0, 0.08135, 0, 0.5949). Summarizing and writing collectively, the strategies in terms of PULC are: ${\varrho_1}^* = (0, 0.7717, 0.3237)$, ${\varrho_2}^* = 0, {\varrho_3}^* = (0.305, 0.2283, 0.0813), {\varrho_4}^* = 0, {\varrho_5}^* = (0.695, 0, 0.5949)$ and the PULC value of the game is $\upsilon^- = (1.36, 2.93, 0.96)$. Further, the probabilistic unbalanced linguistic value of the game can be calculated and is equal to $\widetilde{v}^-(p) = \{t_{2^0}(0.4612), t_{2^1}(0.5388)\}$.

Similarly, solving the game problem for Player *II*, i.e. for "Year", the expectation, entropy and hyperentropy games are as given below: Expectation Game

$$\min \quad {\upsilon_{EX}}^+$$
$$\text{s.t.} \begin{cases} 0.6298\varphi_1 + 0.9399\varphi_2 + 1.0901\varphi_3 + 1.6112\varphi_4 + 2.6473\varphi_5 + 0.3765\varphi_6 \le {\upsilon_{EX}}^+ \\ 0\varphi_1 + 0\varphi_2 + 0\varphi_3 + 0.5717\varphi_4 + 1.9864\varphi_5 + 0.5717\varphi_6 \le {\upsilon_{EX}}^+ \\ 2.3255\varphi_1 + 3.2944\varphi_2 + 0.6284\varphi_3 + 1.2569\varphi_4 + 2.5139\varphi_5 + 2.5139\varphi_6 \le {\upsilon_{EX}}^+ \\ 0.6056\varphi_1 + 0.8091\varphi_2 + 1.0077\varphi_3 + 0.3405\varphi_4 + 2.4252\varphi_5 + 1.387\varphi_6 \le {\upsilon_{EX}}^+ \\ 0.9468\varphi_1 + 1.3413\varphi_2 + 1.6943\varphi_3 + 1.7911\varphi_4 + 4.4936\varphi_5 + 2.6826\varphi_6 \le {\upsilon_{EX}}^+ \\ \text{where } \varphi_j \ge 0,\ \sum_{j=1}^{6} \varphi_j = 1 \end{cases}$$

Solving by doing required substitutions, we get the optimal solution as: ${\upsilon_{EX}}^+ = 1.36$, and the strategies are (0.4359, 0, 0.564, 0, 0, 0). Entropy Game

$$\max \quad {\upsilon_{EN}}^{+}$$
$$\text{s.t.} \begin{cases} \sqrt{(1.7603)^2\varphi_1 + (2.6269)^2\varphi_2 + (3.0467)^2\varphi_3 + (2.2799)^2\varphi_4 + (3.0467)^2\varphi_5 + (0.4333)^2\varphi_6} \geq {\upsilon_{EN}}^{+} \\ \sqrt{(1.854)^2\varphi_1 + (1.854)^2\varphi_2 + (1.854)^2\varphi_3 + (1.5978)^2\varphi_4 + (5.5518)^2\varphi_5 + (1.5978)^2\varphi_6} \geq {\upsilon_{EN}}^{+} \\ \sqrt{(3.2907)^2\varphi_1 + (3.7914)^2\varphi_2 + (0.6284)^2\varphi_3 + (1.2569)^2\varphi_4 + (2.5139)^2\varphi_5 + (2.5139)^2\varphi_6} \geq {\upsilon_{EN}}^{+} \\ \sqrt{(1.6926)^2\varphi_1 + (2.2613)^2\varphi_2 + (2.8165)^2\varphi_3 + (0.482)^2\varphi_4 + (3.4317)^2\varphi_5 + (1.9626)^2\varphi_6} \geq {\upsilon_{EN}}^{+} \\ \sqrt{(1.3397)^2\varphi_1 + (1.5436)^2\varphi_2 + (1.9499)^2\varphi_3 + (1.7911)^2\varphi_4 + (4.4936)^2\varphi_5 + (3.0873)^2\varphi_6} \geq {\upsilon_{EN}}^{+} \\ \text{where } \varphi_j \geq 0,\ j = 1, \ldots, 6\ \sum_{j=1}^{6} \varphi_j = 1 \end{cases}$$

Solving just as for Player *I*, we obtain result for entropy game as (0, 0.2839, 0, 0, 0.716, 0) and value ${\upsilon_{EN}}^{+} = 2.93$. Similarly for hyperentropy, Hyper-entropy Game

$$\max \quad {\upsilon_{HE}}^{+}$$
$$\text{s.t.} \begin{cases} \sqrt{(0.6987)^2\varphi_1 + (1.0427)^2\varphi_2 + (1.2093)^2\varphi_3 + (1.427)^2\varphi_4 + (1.2093)^2\varphi_5 + (0.172)^2\varphi_6} \geq {\upsilon_{HE}}^{+} \\ \sqrt{(0.3506)^2\varphi_1 + (0.3506)^2\varphi_2 + (0.3506)^2\varphi_3 + (0.6342)^2\varphi_4 + (2.203)^2\varphi_5 + (0.6342)^2\varphi_6} \geq {\upsilon_{HE}}^{+} \\ \sqrt{(2.0596)^2\varphi_1 + (1.505)^2\varphi_2 + (0.1188)^2\varphi_3 + (0.3015)^2\varphi_4 + (2.1479)^2\varphi_5 + (1.2284)^2\varphi_6} \geq {\upsilon_{HE}}^{+} \\ \sqrt{(0.6718)^2\varphi_1 + (0.8976)^2\varphi_2 + (1.118)^2\varphi_3 + (0.3015)^2\varphi_4 + (2.1479)^2\varphi_5 + (1.2284)^2\varphi_6} \geq {\upsilon_{HE}}^{+} \\ \sqrt{(0.8385)^2\varphi_1 + (0.6127)^2\varphi_2 + (0.774)^2\varphi_3 + (0.3387)^2\varphi_4 + (0.8497)^2\varphi_5 + (1.2255)^2\varphi_6} \geq {\upsilon_{HE}}^{+} \\ \text{where } \varphi_j \geq 0,\ j = 1, \ldots, 6\ \sum_{j=1}^{6} \varphi_j = 1 \end{cases}$$

The solution comes out as: (0.173, 0, 0, 0, 0.5669, 0.2598) and ${\upsilon_{HE}}^{+} = 0.96$. Hence, collectively the optimal PULC strategies and PULC value of Player *II* are: ${\varphi_1}^* = (0.4359, 0, 0.173)$, ${\varphi_2}^* = (0, 0.2839, 0)$, ${\varphi_3}^* = (0.564, 0, 0)$, ${\varphi_4}^* = 0$, ${\varphi_5}^* = (0, 0.716, 0.5669)$, ${\varphi_6}^* = (0, 0, 0.2598)$ Value $\upsilon^{+} = (1.36, 2.93, 0.96)$. Hence, $\upsilon^{-} = \upsilon^{+} = \upsilon =$ Value of the game. Further, the scores of strategies may be evaluated to finalize which strategy should a player use to win the game. For that, we may find different score function values for different values of γ. Without loss of generality, we fix the value of $\gamma = 0.5$ for current calculations.

For Player *I*, $\hat{S}({\varrho_1}^*) = -0.4184$, $\hat{S}({\varrho_2}^*) = 0$, $\hat{S}({\varrho_3}^*) = 0.3825$, $\hat{S}({\varrho_4}^*) = 0$, $\hat{S}({\varrho_5}^*) = 0.3975$. Thus, based on the score values, the IT company Larsen & Toubro would be the best to invest the money in. Thus, based on the above-obtained results, the DM can also plan where to invest in the future. Also, the trend of increment of stock prices may be determined by finding the score values of Player *II*, i.e. "Year".

Note. The above problem is solved for the case when the DMs record only a single linguistic term as a payoff entry. The same procedure may be followed for the case when multiple linguistic terms consist a payoff or the case when the payoffs are of a set type. In that case, we could break the individual term of the set and then form separate PULCLP for each broken term set.

5 Analysis and discussion

This section includes the analysis of numerical results obtained in previous section along with the comparison of proposed approach with the formerly defined literature.

5.1 Numerical discussion

For the numerical example 3, the matrix $\widetilde{A}$ corresponds to the first player, i.e. "IT Company" and based on the calculations, the DM is supposed to analyse which IT company is the best to invest in. For this purpose, the original problem is divided into three individual games, namely expectation, entropy and hyperentropy game. Solving the three games, it can be concluded that the optimal strategies are $\varrho_1{}^* = (0, 0.7717, 0.3237), \varrho_2{}^* = 0, \varrho_3{}^* = (0.305, 0.2283, 0.0813), \varrho_4{}^* = 0, \varrho_5{}^* = (0.695, 0, 0.5949)$ and the PULC value of the game is $\upsilon^- = (1.36, 2.93, 0.96)$. The score of the strategies may be evaluated as $\hat{S}(\varrho_1{}^*) = -0.4184, \hat{S}(\varrho_2{}^*) = 0, \hat{S}(\varrho_3{}^*) = 0.3825, \hat{S}(\varrho_3{}^*) = 0.3975$. Thus, based on the score values, the IT company Larsen & Toubro would be the best to invest the money in.

Similar task can be done for the second player, i.e. "Year" that can be used to investigate which year was the best year for investment in a particular IT company. For that purpose, similar expectation, entropy and hyperentropy game problems are constituted and solved.

5.2 Comparative analysis and discussion

This section provides a comparison analysis of the proposed methodology with the existing studies that involve game theoretic models using PULCs.

The idea of probabilistic linguistic term set (PLTS) was first proposed by Pang et al. [31], which allows the players or DMs to evaluate situations more effectively by using multiple terms rather than just one, by giving each term a probability-based weight. Apart from this, there are other variants of PLTS, such as [3, 46, 51]. In the paper [3], the authors have defined a concept "Interval-valued probabilistic linguistic term sets" that can describe two attributes of an object: possible linguistic term sets and their associated interval numbers. The former provides an assessment, such as "perfect" or "low". The latter describes the hesitancy for the given assessment and denotes the interval-valued probabilistic distribution or membership degrees with respect to the former. Xie et al. [46] extended the analytic hierarchy process to the situations with the probabilistic linguistic information. Zhai et al. [51] focus on improving the accuracy of the multi-granular linguistic recommender system by supporting customers to find out the most eligible items according their own preferences. For this purpose, they propose a probabilistic linguistic vector-term sets to promote the application of multi-granular linguistic information. However, despite having several practical applications, these PLTSs have the limitation that all of them use symmetric and balanced information for decision analysis. Further, this

symmetric information inculcated in these PLTSs is also used in matrix game theory in the paper [30]. The theory defined by Mi et al. [30] uses triangular membership function to represent probabilistic linguistic information. This information is defuzzified using the triangular membership function, and the calculated games values are then retranslated to probabilistic data by using linear interpolation method. This process is limited in its form as the defuzzification, and then, retranslation process causes major loss of information. Also, Mi et al. [30] study the cloud model in game problems where the responses of players belong to a symmetrically distributed linguistic set. To overcome the problem of symmetric probabilistic linguistic information, Han et al. [16] defined a probabilistic unbalanced linguistic term set (PULTS) that makes use of unbalanced linguistic data to represent probabilistic linguistic information. But this method uses the unbalanced linguistic model defined by [19] which is computationally very expensive method. Thus, in the paper [27], the authors solve linguistic matrix games that have probabilistic unbalanced linguistic data that comes from a probabilistic multiplicative unbalanced linguistic set defined by Malhotra & Gupta [28]. Table 3 summarizes the above-mentioned comparison analysis in tabular form.

The theory defined in our paper is useful for solving matrix games having the same probabilistic multiplicative linguistic information, but with a different point of view. The authors in [27] only considered the fuzziness involved in a linguistic concept. However, in our paper, the linguistic concept is analysed in the way that both randomness and fuzziness are omitted from the qualitative concept. The present paper uses cloud model [11] to thoroughly capture the intrinsic fuzziness and vagueness of decision information present in matrix game problems. Apart from that, the game problem considered in this paper consists of payoffs that are randomly irregularly distributed in a multiplicative manner. Such linguistic data address all types of linguistic information. This preserves the literal meaning of unbalanced linguistic matrix games. The suggested approach not only prevents data loss, it also studies different types of uncertainties lying in the linguistic information of matrix game problems. Also, the loss of information occurring in [30] due to the retranslation process is well handled in this paper as the probabilistic information is first converted into PULCs and the PULC data are then solved using cloud aggregation operators. Also, the retranslation process in this paper is done using the bijective operators that overcome the information loss. Furthermore, the methodology defined in this paper converts the probabilistic unbalanced payoff into a PULC in a way that it forms three independent easy-to-solve games, namely Expectation game, Entropy game and Hyperentropy game, which are basically three independent PULLPs. These PULLPs are converted into LPPs using required substitution and are then solved as a classical LPP. The objective function of these PULLPs runs into $\mathcal{O}(n)$ time, where n is the cardinality of the PULTS.

The novelty of this paper lies in the fact that it allows the DMs to study the underlying uncertainty in the information of an uncertain game problem and still solve a decision-making problem just as a classical problem without any loss of information. Unlike existing matrix game problems, this study captures the intrinsic randomness and fuzziness present in a qualitative concept of a matrix game which is based on a PM-ULTS.

Table 3 Comparison of the defined methodology with some existing research

Paper	Nature of information	Distribution of linguistic terms	Application in game	Nature of uncertainty of payoffs
[31]	Probabilistic	Symmetric	No	N/A
[3]	Probabilistic	Symmetric	No	N/A
[46]	Probabilistic	Symmetric and Asymmetric	No	N/A
[30]	Probabilistic	Symmetric	Yes	Fuzziness
[16]	Probabilistic	Asymmetric	No	N/A
[27]	Probabilistic	Asymmetric	Yes	Fuzziness
This paper	Probabilistic	Asymmetric	Yes	Randomness & Fuzziness

6 Conclusion

This study proposed and investigated the two-person matrix games having probabilistic multiplicative unbalanced linguistic information. This allows the theory of games to accept incomplete information as players' response as an input. The uncertain payoffs are dealt with by PULCs, and then, cloud aggregation operators are applied to aggregate the probabilistic cloud information. Then, the game problem is converted into three independent games, i.e. expectation game, entropy game and hyperentropy game. The LPPs of each game are individually solved to get cloud value of the game. The significance of transforming the unbalanced linguistic payoffs to PULCs is that the cloud model studies the uncertainty lying in a qualitative concept by thoroughly encapsulating the underlying randomness and fuzziness of the unbalanced payoff. The results of all three LPPs are collected to form a single strategy and single value, that further form a cloud value. The cloud value of the game provides the dispersion of cloud droplets using entropy value. Further, the hyperentropy studies the uncertainty of entropy. Thus, this methodology is advantageous as it provides more accurate results than other existing methodologies to solve matrix game problems having probabilistic linguistic information, in which only one kind of uncertainty is dealt. The more uncertainty is dealt, the more we proceed towards the preciseness of a certain result. Afterwards, this methodology is applied to solve a real-time decision problem of investment in stock of different IT companies.

As for future research directions, one can extend this methodology towards interval clouds as this methodology maps a qualitative information to a certain precise value. On the other hand, interval provides a more generalized range, thus reducing the loss of information. Further, one can work towards developing a more simpler theory for the case where the payoff information is in the form of multiple probabilistic linguistic terms.

Author contributions Parul Chauhan contributed to conceptualization, methodology, writing and editing, and Anjana Gupta contributed to methodology, formal analysis and review.

Funding The authors did not receive support from any organization for the submitted work.

Data availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Code availability Not applicable.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Arfi B (2006) Linguistic fuzzy-logic game theory. J Confl Resolut 50(1):28–57
2. Arfi B (2006) Linguistic fuzzy-logic social game of cooperation. Ration Soc 18:471–537
3. Bai C, Zhang R, Shen S, Huang C, Fan X (2018) Interval-valued probabilistic linguistic term sets in multi-criteria group decision making. Int J Intell Syst 33(6):1301–1321. https://doi.org/10.1002/int.21983
4. Bao G, Lian X, He M, Wang L (2010) Improved two-tuple linguistic representation model based on new linguistic evaluation scale. Control Decis 25(5):780–784
5. Barron EN (2009) Game theory: an introduction. Wiley, New Jersy
6. Bartczuk L, Dziwínski P, Starczewski JT (2012) A New Method for Dealing with Unbalanced Linguistic Term Set. ICAISC 2012. Part I, LNCS 7267:207–212
7. Bector CR, Chandra S, Vidyottama V (2004) Matrix games with fuzzy goals and fuzzy linear programming duality. Fuzzy Optim Decis Making 3:255–269
8. Butnariu D (1978) Fuzzy games: a description of the concept. Fuzzy Set Syst 1:181–92
9. Cai M, Gong Z, Yu X (2016) A method for unbalanced linguistic term sets and its application in group decision making. Int J Fuzzy Syst 19(3):671–682
10. Campos L (1989) Fuzzy linear programming models to solve fuzzy matrix games. Fuzzy Sets Syst 32:275–289
11. Chauhan P, Gupta A, Malhotra T (2023) A novel cloud model based on a multiplicative unbalanced linguistic term set. J Supercomput. https://doi.org/10.1007/s11227-023-05295-z
12. Chauhan P, Gupta A (2021) Matrix games with proportional linguistic payoffs. Soft Comput 25:15067–15081
13. Chauhan P, Gupta A (2022) Matrix games with linguistic distribution assessment payoffs. In: Saraswat M, Sharma H, Balachandran K, Kim JH, Bansal JC (eds) Congress on intelligent systems. Lecture notes on data engineering and communications technologies. Springer, Singapore
14. Dengfeng L (1999) A fuzzy multiobjective programming approach to solve fuzzy matrix games. J Fuzzy Math 7(4):907–912
15. Dong Y, Xu Y, Yu S (2009) Computing the numerical scale of the linguistic term set for the 2-tuple fuzzy linguistic. IEEE Trans Fuzzy Syst 17(6):1366–1378
16. Han B, Tao Z, Chen H, Zhou L, Liu J (2020) A new computational model based on Archimedean copula for probabilistic unbalanced linguistic term set and its application to multiple attribute group decision making. Comput Ind Eng 140:106–264
17. Herrera F, Martínez L (2000) A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Trans Fuzzy Syst 8(6):746–752

18. Herrera F, Martínez L (2001) The 2-tuple linguistic computational model, advantages of its linguistic description, accuracy and consistency. Int J Uncertain Fuzziness Knowl Based Syst 9:33–48
19. Herrera F, Herrera-Viedma E, Martínez L (2008) A fuzzy linguistic methodology to deal with unbalanced linguistic term sets. IEEE Trans Fuzzy Syst 16(2):354–370
20. Jin C, Wang H, Xu Z (2019) Uncertain probabilistic linguistic term sets in group decision making. Int J Fuzzy Syst 21:1241–1258
21. Li D, Liu C, Gan W (2009) A new cognitive model: cloud model. Int J Intell Syst 24:357–375
22. Li DF (2013) An effective methodology for solving matrix games with fuzzy payoffs. IEEE Trans Cybern 43(2):610–621
23. Li LL, Liu LX, Yang CW, Li ZG (2012) The comprehensive evaluation of smart distribution grid based on cloud model. Energy Proc 17:96–102
24. Liao LC, Guo WW (2012) Incorporating utility and cloud theories for owner evaluation in tendering. Expert Syst Appl 39:5894–5899
25. Liu ST, Kao C (2007) Solution of fuzzy matrix games: an application of the extension principle. Int J Intell Syst 22:891–903
26. Liu W, Zhu Z (2021) A multistage decision-making method for multi-source information with Shapley optimization based on normal cloud models. Appl Soft Comput 111:107716
27. Malhotra T, Gupta A (2023) Probabilistic multiplicative unbalanced linguistic term set and its application in matrix games. Int J Mach Learn Cybern 14:1253–1283
28. Malhotra T, Gupta A (2020) A new 2-tuple linguistic approach for unbalanced linguistic term sets. IEEE Trans Fuzzy Syst 29:2158–2168
29. Meng XY, Zhang GW, Kang JC, Li HS, Li DY (2008) A new subjective trust model based on cloud model. IEEE International Conference on Networking, Sensing and Control, ICNSC '08:1125–1130
30. Mi X, Liao H, Zeng XJ, Xu Z (2021) The two-person and zero-sum matrix game with probabilistic linguistic information. Inf Sci 570:487–499
31. Pang Q, Wang H, Xu Z (2016) Probabilistic linguistic term sets in multi-attribute group decision making. Inf Sci 369:128–143
32. Sakawa M, Nishizaki I (1994) Max-min solutions for fuzzy multiobjective matrix games. Fuzzy Sets Syst 67:53–69
33. Singh A, Gupta A (2018) Matrix games with interval-valued 2-tuple linguistic information. Games 9(3):62
34. Singh A, Gupta A, Mehra A (2018) Matrix games with 2-tuple linguistic information. Ann Oper Res 287:895–910
35. Tang G, Yang Y, Gu X et al (2022) A new integrated multi-attribute decision-making approach for mobile medical app evaluation under q-rung orthopair fuzzy environment. Expert Syst Appl 200:117034
36. Tang G, Long J, Gu X et al (2022) Interval type-2 fuzzy programming method for risky multicriteria decision-making with heterogeneous relationship. Inf Sci 584:184–211
37. Tang G, Zhang X, Zhu B et al (2023) A mathematical programming method based on prospect theory for online physician selection under an R-set environment. Inf Fusion 93:441–468
38. Tang G, Gu X, Chiclana F et al (2023) A multi-objective q-rung orthopair fuzzy programming approach to heterogeneous group decision making. Inf Sci 645:119343
39. Wang P, Huang S, Cai C (2021) Dual linguistic term set and its application based on the normal cloud model. IEEE Trans Fuzzy Syst 29(8):2180–2194
40. Wang P, Xu X, Huang S, Cai C (2018) A linguistic large group decision making method based on the cloud model. IEEE Trans Fuzzy Syst 26(6):3314–3326
41. Wang HL, Feng YQ (2005) On multiple attribute group decision making with linguistic assessment information based on cloud model. Control Decis 20(6):679–685
42. Wang JQ, Peng L, Zhang HY, Chen XH (2014) Method of multi-criteria group decision-making based on cloud aggregation operators with linguistic information. Inf Sci 274:177–191
43. Wang X, Wang Y, Zhang H, Wang J, Li L, Goh M (2021) An asymmetric trapezoidal cloud-based linguistic group decision-making method under unbalanced linguistic distribution assessments. Comput Ind Eng 160:107457
44. Wang JH, Hao J (2006) A new version of 2-tuple fuzzy linguistic representation model for computing with words. IEEE Trans Fuzzy Syst 14(3):435–445

45. Wu Y, Zhang Z, Kou G, Zhang H et al (2021) Distributed linguistic representations in decision making: taxonomy, key elements and applications, and challenges in data science and explainable artificial intelligence. Inf Fusion 65:165–178
46. Xie W, Xu Z, Ren Z, Wang H (2018) Probabilistic linguistic analytic hierarchy process and its application on the performance assessment of Xiongan new area. Int J Inf Technol Decis Making. https://doi.org/10.1142/s0219622017500420
47. Yi Z (2022) Decision-making based on probabilistic linguistic term sets without loss of information. Complex Intell Syst 8:2435–2449
48. Yue N, Xie J, Chen S (2020) Some new basic operations of probabilistic linguistic term sets and their application in multi-criteria decision making. Soft Comput 24:12131–12148
49. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
50. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning, I-III. Inf Sci 8:301–357
51. Zhai Y, Xu Z, Liao H (2016) Probabilistic linguistic vector-term set and its application in group decision making with multi-granular linguistic information. Appl Soft Comput 49:801–816
52. Zhang G, Dong Y, Xu Y (2014) Consistency and consensus measures for linguistic preference relations based on distribution assessments. Inf Fusion 17:46–55
53. Zhang Z, Yu W, Martinez L, Gao Y (2019) Managing multigranular unbalanced hesitant fuzzy linguistic information in multiattribute large-scale group decision making: a linguistic distribution-based approach. IEEE Trans Fuzzy Syst 28(11):2875–2889
54. Zhang Z, Li Z, Gao Y (2021) Consensus reaching for group decision making with multi-granular unbalanced linguistic information: A bounded confidence and minimum adjustment-based approach. Inf Fusion 74:96–110
55. Zhang Z, Li Z (2022) Personalized individual semantics-based consistency control and consensus reaching in linguistic group decision making. IEEE Trans Syst Man Cybern Syst 52(9):5623–5635

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Check for updates

Received: 14 February 2024 | Revised: 21 April 2024 | Accepted: 1 May 2024

DOI: 10.1002/ctpp.202400020

ORIGINAL ARTICLE

Contributions to Plasma Physics

Radially polarized femtosecond laser interaction with unmagnetized plasma slab and symmetric modes for enhanced terahertz field generation

Himank Sagar | **Suresh C. Sharma**

Department of Applied Physics, Delhi Technological University, Delhi, India

Correspondence
Suresh C. Sharma, Department of Applied Physics, Delhi Technological University, Shahbad Daulatpur, Bawana Road, Delhi 110 042, India.
Email: suresh321sharma@gmail.com

Abstract

We consider the excitation of terahertz (THz) electromagnetic fields by interaction of radially polarized laser pulses of diverse profiles with a homogenous plasma density slab. We utilize the properties of the laser pulse to generate THz fields in a plasma slab. It is shown that the radial ponderomotive force exerted by laser imparts an oscillatory velocity to plasma electrons and drives a nonlinear current in azimuthal direction exciting THz electromagnetic fields in the plasma slab. The dependence of the excited radial electric field and azimuthal magnetic field on axial and radial parameters of the plasmas lab, as well as on the slab thickness and laser pulse width size, is investigated. It is demonstrated that the terahertz fields are generated most efficiently with a frequency close to the plasma frequency. It is also shown that the intensity of the excited fields may be optimized and controlled by the plasma slab and laser pulse parameters. Rectangular-triangular, super-Gaussian, and sinusoidal lasers exhibit a significantly steeper radial gradient of ponderomotive potential in comparison with other laser profiles, and excite intense radial electric fields and generate azimuthal magnetic fields in plasma slab. The numerical results closely follow the scaling laws and match with previous experimental and simulation results.

KEYWORDS

beam width, electromagnetic field, laser profile, ponderomotive potential, radial polarization

1 | INTRODUCTION

Laser-plasma interaction is an important process in the generation of electromagnetic fields and relativistic particles. The process is driven by the interaction of high-intensity laser pulses with a plasma. The laser energy creates an electric current in the plasma, which in turns generates a magnetic field. Recent studies have shown that the laser-plasma interaction can be enhanced by using a thin plasma slab for better control and improved absorption of the laser beam.[1,2] When the laser frequency exceeds the plasma frequency, the plasma becomes transparent, as the laser photons are able to travel through the plasma without being absorbed. As a result, the tunneling of the laser energy may take place through the plasma slab. Intense magnetic fields of several megagauss (MG) are self-generated in this regime of interaction. Such magnetic fields in plasma find numerous applications in magnetic field reconnection research,[3,4] strong electron acceleration,[5] fast ignition transport studies,[6,7] generation of shock waves in collision-less plasmas,[8] etc. The generation of terahertz (THz) electric fields using laser-induced plasma slabs is a well-established technique in the field of ultrafast optics and

 © 2024 Wiley-VCH GmbH

THz spectroscopy. This method relies on the interaction of intense femtosecond laser pulses with a gas or solid target, which generates a plasma that can emit or modulate THz radiation. Electromagnetic terahertz (THz) radiations are utilized in many powerful applications, such as in imaging,[9] biomedical,[10] food processing,[11–13] communication,[14] development of future generation devices,[15,16] and military applications.[17]

The state of laser's polarization impacts the optical behavior, focus, and reflectance of laser. A radically polarized laser can accelerate electrons to relativistic energies by trapping the electrons via azimuthal magnetic field and radial electric field components and accelerating them via axial component of electric field. The electromagnetic fields are excited inside and outside the plasma slab by the ponderomotive forces of laser radiation. The shape and polarization of the laser pulse affects the efficiency of THz emission. An oscillatory velocity to the plasma electrons is exerted by the nonlinear ponderomotive force. The coupling of the velocity at the plasma-vacuum interface produces a rotational current, which is primarily responsible to excite electromagnetic radiation, which in turn produces waves of wakefield oscillations.[18–20] The influence of this ponderomotive force on plasma behavior has been extensively explored in a comprehensive review authored by Lundin and Guglielmi.[21] Their analysis delves into how this force is affected by several key factors, including wave amplitude, variations in the ambient magnetic field, wave non-stationarity, the specific wave mode employed, and various other parameters related to both the plasma and the wave. Electric fields in the THz range can be excited in radial and axial directions by interaction of a linearly polarized (s-polarized, p-polarized), circularly polarized or radially polarized laser pulse with plasma.[22,23] However, magnetic fields can be exerted only in an inhomogeneous plasma or with asymmetric laser pulses. The intensity of the laser can become inhomogeneous due to small plasma gradient, self-modulation of pulse wavefront, laser backscattering, or plasma refractive index disparity.[24]

Recently, several authors have performed investigations on self-generated magnetic fields driven by a laser in plasma experimentally[25–27] as well as theoretically.[28–31] An axial magnetic field in Megagauss (MG) range has been observed due to inverse Faraday effect from relativistic laser interactions with an underdense plasma. Scientists have made substantial advancements in the field of laser-plasma interactions for creating broadband THz sources in theoretical[32–42] as well as experimental[43–58] researches. Malik et al.[59] passed a triangle laser beam across a plasma with a ripple density and an external magnetic field and were able to produce a tunable THz source. THz radiation resulted from this interaction, highlighting the significance of perfect plasma management. Bakhtiari et al.[60] studied to maximize the efficiency of THz generation for real-world applications and their study methodically investigated variables like laser intensity and plasma density. By utilizing phase matching techniques and optical mixing with two-color lasers in a plasma medium with ripple density, Kumar and Tripathi were able to generate resonant THz waves.[61] The development of THz sources for spectroscopy, imaging, and communications is advanced by this work. In conclusion, these contributions advance the development of THz sources by accurate plasma regulation, parameter optimization, and cutting-edge methods with prospective applications in numerous sectors. The terahertz radiation emission from laser-irradiated objects inside a typically inhomogeneous plasma has been thoroughly investigated by the authors. Their investigation demonstrates that this emission is caused by linear mode conversion, offering an important explanation for this puzzling phenomenon.

In our continuing research, we analyze the establishment of radial and axial THz electric fields and self-generated magnetic field in an homogeneous plasma using the ponderomotive force based method. This force, which outcomes from the spatial imperfections in the laser pulse, is essential for producing an azimuthal magnetic field during interactions with homogeneous plasma. The objective is to explore novel methods for generating terahertz (THz) radiations at the interface between plasma and vacuum environment. This investigation delves into the intricate realm of laser beam shapes, specially sinusoidal, Gaussian, cosh-hyperbolic, super-Gaussian, triangular, and rectangular-triangular beams deployed both within the plasma slab and in the vacuum region. Our approach hinges on the meticulous application of continuity equations and boundary conditions at the critical juncture where plasma and vacuum converge. Within this framework, we model the intricate interaction between the diverse laser beam shapes and the plasma slab, delving into wave equations, nonlinear plasma responses and the perturbations these beams induce in the electron density. Our ultimate aim is to decipher the mechanisms underlying THz radiations generation stemming from these laser-plasma interactions. This investigation further extends to the magnetic fields produced by these distinct profiles, also presenting how the viewing angle affects the resultant model. By means of advanced simulations, we construct three-dimensional waveforms that delineate intricate correlations within the radial and axial dimensions, in conjunction with the electric and magnetic fields arising from diverse laser profiles, both within the confines of plasma and in vacuum. We meticulously gauge the intensities of these magnetic and electric fields, undertaking a rigorous evaluation of their respective THz field strengths. Furthermore, "surf" plots serve as instrumental tools for the graphical representation and in-depth scrutiny of the disparate behaviors exhibited by various waveforms when immersed in a uniform plasma environment. The authors have also undertaken research pertaining to flattened laser beams, with a primary objective of quantifying the generation

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of THz radiation through the phenomenon known as phase matching resonance.[62] The radial direction of electric field and the azimuthal direction of magnetic field are focused on in both 2D and 3D contour plots to investigate the intensity of corresponding fields using simulations. By comparing the observed results with the provided experimental and theoretical framework, the concept is convincingly illustrated. The final section of this article offers a brief summary of the research findings.

2 | MATHEMATICAL ANALYSIS

We propose a technique for producing THz field in the range of MV/cm by the interaction of laser with the plasma slab. We have taken into account a moderate density plasma slab of nonlinear material of width 2 L in the z-direction with its dimensions ranging from $-L < z < L$. Vacuum regions are selected along both sides of the plasma slab. The presence of a vacuum-plasma interface is an established and viable phenomenon within the realm of plasma physics, a field where high-energy electron beams or laser pulses intersect with plasmas. This boundary between the vacuum and the plasma is fundamental and plays a pivotal role in numerous plasma physics experiments and practical applications that are still used in many research areas. When laser beam interacts the slab boundary, it excites the radial E_r and axial E_z components of electric field and azimuthal B_ϕ component of magnetic field. Various laser beam profiles have been studied in this paper for these constituents of fields.

The density of the plasma slab is rendered in such a way that the electron number density remains constant within the slab along the z-direction. The density is a constant step function whose value is constant in the plasma, that is, $-L\langle z\langle L$ and is 0 outside the slab, that is, $z \leq -L \& z \geq L$ (in vacuum). We visualize the electron plasma density as μ. In the current scenario, the pulse propagates through the plasma slab without being fully transmitted. This is due to the fact because $\omega_o/\omega_p > 1$. When the laser's frequency surpasses that of the plasma, a remarkable phenomenon occurs—the plasma becomes translucent, allowing laser photons to traverse it without being absorbed. Consequently, the laser's energy can tunnel through the plasma slab, offering intriguing possibilities in various scientific and practical domains, such as laser-driven fusion and advanced particle acceleration. The general equation of electric field for the input laser profile is a function of r, z, and t which can be written as

$$E_{\text{input}} = E_{\text{ol}} e^{(-i\omega_o t + ik_z z)} f(r)_{\text{profile}}, \tag{1}$$

where $E_{\text{ol}} = 10^7\ \text{V cm}^{-1}$ is the maximum laser amplitude, $\omega_o = 2.0 \times 10^{14}\text{Hz}$ is the incident laser frequency, and $k_z = \frac{\omega_o}{c}\text{cm}^{-1}$ is the wave number of laser beam in the z-direction.

In this study, we have investigated a range of laser radial profiles, including Gaussian, sinusoidal, cosh-hyperbolic, super-Gaussian, triangular, and rectangular-triangular. Each of these profiles introduce variations in the function denoted as $f(r)$, which in turn influences the parameter $k_\perp$ specific to each profile. The laser beam imparts an oscillatory velocity to electrons and exerts a nonlinear ponderomotive force on them. The interaction of electron velocity with sharp density gradients, particularly at the plasma-vacuum interface, gives rise to a rotational current. Consequently, the abrupt density changes at the interfaces between the vacuum and plasma play a pivotal role in exciting electromagnetic radiation. The ponderomotive force acting on plasma electrons is defined by a mathematical equation given as

$$\varphi = \frac{eE_{\text{ol}}^2 |f(r)|^2}{4m_e\omega_o^2}, \tag{2}$$

where $e = 4.8032 \times 10^{-10}$ statcoulomb is the charge of electron and $m_e = 9.109 \times 10^{-28}\text{g}$ is the mass of electron.

The transverse wave number, abbreviated as "$k_\perp$" is not steady but changes depending on the profile and adjusts as the input laser beam's shape alters. Below are listed the wave numbers for the ponderomotive potential at various radial distances "r" for the distinct profiles under study.

For sinusoidal laser beam:

$$f(r) = \sin\left(\frac{\pi r}{b_w}\right), \tag{3}$$

$$k_\perp \varphi = 2E_{\text{ol}}^2 f(r)\left(\frac{e^2}{4m_e\omega_o^2}\right)\left(\frac{\pi}{b_w}\right)\cos\left(\frac{\pi r}{b_w}\right). \tag{4}$$

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

For Gaussian and Super-Gaussian laser beam:

$$f(r) = e^{-\left(\frac{r^2}{b_w^2}\right)^n}, \tag{5}$$

$$k_\perp \varphi = \frac{-4nr}{b_w^2}\left(\frac{eE_{ol}^2}{4m_e\omega_o^2}\right)e^{-2\left(\frac{r^2}{b_w^2}\right)^n}\left(\frac{r^2}{b_w^2}\right)^{n-1}. \tag{6}$$

For triangular and rectangular-triangular laser beam,

$$f(r) = 1 - \left|\frac{r}{b_w}\right|^s, \tag{7}$$

$$k_\perp \varphi = 2\left[1 - \left|\frac{r}{b_w}\right|^s\right]\left(\frac{eE_{ol}^2}{4m_e\omega_o^2}\right)\left(\frac{-s}{b_w}\right)\left|\frac{r}{b_w}\right|^{s-1}. \tag{8}$$

For cosh-hyperbolic laser beam:

$$f(r) = \cosh\left(\frac{br}{b_w}\right)e^{\left(-\frac{r^2}{b_w^2}\right)}, \tag{9}$$

$$k_\perp \varphi = \left(\frac{eE_{ol}^2}{2m_e\omega_o^2}\right)e^{\left(-\frac{r^2}{b_w^2}\right)}\left[\left(\frac{-2r}{b_w^2}\right)\cosh^2\left(\frac{br}{b_w}\right) + \left(\frac{b}{2b_w}\right)\sinh\left(\frac{2br}{b_w}\right)\right]. \tag{10}$$

Within the aforementioned equations, the variable "n" serves as the form factor. The curvature of the curve intensifies as the value of "n" rises. Precisely, when "n" equals 1, the waveform adopts a Gaussian profile, while for "n" greater than 1 (such as 7), the waveform evolves into a super-Gaussian shape. These adjustments in the "n" parameter hold substantial sway over the waveform's attributes and overall appearance. The parameter "s" represents the order or index of the laser beam. As "s" increases in value, the curve exhibits a greater degree of steepness. To provide specifics, when s equals 1, the waveform assumes a triangular shape, while for values of s such as 4 (typically when s exceeds 2), the waveform transforms into a rectangular-triangular pattern. These variations in "s" have a substantial impact on the shape and characteristics of the laser beam. The parameter "b" serves as a measure of the degree of decentering, and its values encompass a range from 0 to 2. In our specific study, we have chosen to focus on two distinct values of "b," namely $b = 1.2$ and $b = 2$. These particular values have been selected because they exhibit significant disparities in their profile shapes, gradients, and slopes, making them particularly relevant for our investigation. The term "b_w" signifies the beam width of the input laser waveform, which spans from an initial value of 0 and extends up to 40 μm.

Terahertz electromagnetic radiation generation can be described by the hydrodynamic equations and Maxwell's equations averaged over low-frequency laser oscillations for electrons in plasma, and can be expressed as:

$$\frac{\partial}{\partial t}\vec{v} = \frac{e}{m_e}\vec{E} - \frac{e}{m_e}(\nabla\varphi), \tag{11}$$

$$\nabla \times \vec{E} = -\frac{1}{c}\frac{\partial}{\partial t}\vec{B}, \quad \left(a_o = \frac{eE_{ol}}{m_e c\omega}\right), \tag{12}$$

$$\nabla \times \vec{B} = \frac{1}{c}\frac{\partial}{\partial t}\vec{E} + \frac{4\pi e}{c}\vec{J}. \tag{13}$$

Here, $E(r,t)$, $B(r,t)$, and $v(r,t)$ are the electric fields, magnetic fields, and the averaged electron velocities, all altering slowly over the laser field period and c is the velocity of light. Here, $\vec{J} = e\mu\vec{v}$ is the electron current density in plasma produced due to varying electric and magnetic field over a period of time in the region $(-L < z < L)$, and $\vec{v}$ is the normalized velocity of electrons.

The time varying electric and magnetic field equations of the input laser beam can be derived from the Maxwell equations using Equations (11–13), which can be delineated as

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$\frac{\partial^2 \vec{E}}{\partial t^2} + \omega_p^2 \vec{E} + c^2(\nabla \times \nabla \times \vec{E}) = \omega_p^2(\nabla\varphi) \tag{14}$$

$$\frac{\partial^2 \vec{B}}{\partial t^2} + \omega_p^2 \vec{B} - c^2\left(\nabla^2 \vec{B}\right) = 4\pi ec(\nabla n) \times \vec{v}, \tag{15}$$

where $\varphi = \left(\frac{eE_{ol}^2|f(r)^2|}{4m_e\omega_o^2}\right)$ is the averaged ponderomotive potential of laser which is applied on plasma electrons, $\omega_p = \sqrt{\frac{4\pi ne^2}{m_e}}$ is the plasma electron frequency, and n is the number density of electrons in plasma slab with respect to z-direction. In Equation (15), the right side of the equation acts as the originating factor for the generation of THz radiation. Substituting $\partial/\partial t$ by $(-i\omega)$ and ∇ by $\left(ik_\dagger\right)$ in Equation (15) for the azimuthal component of magnetic field, we get

$$(-i\omega)^2 \vec{B}_\phi + \omega_p^2 \vec{B}_\varphi - \left(c^2 \frac{\partial^2}{\partial r^2} \vec{B}_\varphi + \frac{\partial^2}{\partial z^2} \vec{B}_\varphi\right) = 4\pi ec(\nabla n) \times \vec{v}, \tag{16}$$

$$\left(\omega_p^2 - \omega^2\right)\vec{B}_\varphi - c^2\left(k_\dagger^2 \vec{B}_\varphi + \frac{\partial^2}{\partial z^2} \vec{B}_\varphi\right) = 4\pi ec(\nabla n) \times \vec{v}, \tag{17}$$

where $k_\dagger^2 = \frac{\omega}{c}\sin(\theta)$ is the perpendicular wave number and "θ" is the angle at which resultant fields are viewed. In order to assess the power of terahertz radiation produced and determine the electric and magnetic field intensities for varying incident laser beam profiles, it becomes imperative to ascertain the precise angle in question.

The radial component of velocity experienced by the electrons in the plasma due to altering fields is obtained from Equation (11) as

$$(-i\omega_o)\vec{v_r} = -\frac{e}{m_e}\vec{E_r} - \frac{e}{m_e}(k_\perp \varphi), \tag{18}$$

$$\vec{v_r} = \frac{ek_\perp}{im_e}\varphi + \frac{e}{im_e\omega_o}\vec{E_r}. \tag{19}$$

In order to calculate the second-order time varying differential equation for B_φ, we first must conclude V_r in terms of B_φ. To calculate E_r in terms of B_φ, we use Equation (13). The electric field of the laser beam is oriented in the radial (r) direction, which produces the magnetic field aligned with azimuthal-axis, which gives rise to the equation:

$$E_r = \frac{c}{i\omega_o}\frac{\partial B_\varphi}{\partial z}. \tag{20}$$

Substituting Equation (20) in Equation (19), we get radial velocity in terms of B_φ which yields Equation (7) as

$$v_r = \frac{ek_\perp}{\omega_o m_e}\varphi + \frac{ec}{m_e\omega_o^2}\frac{\partial}{\partial z}B_\varphi. \tag{21}$$

The n, that is, the number density varies along the axial direction while the v_r as discussed in Equation (21) varies along the radial direction, which yields the result along azimuthal direction stated as

$$-\omega^2 \vec{B}_\phi + \omega_p^2 \vec{B}_\varphi - c^2\left(k_\dagger^2 \vec{B}_\varphi + \frac{\partial^2}{\partial z^2} \vec{B}_\varphi\right) = \left(\frac{4\pi e^2 c\vec{k_\perp}\varphi}{\omega_o m_e}\right) - \left(\frac{4\pi ec^2}{\omega_o^2 m_e}\frac{\partial n}{\partial z}\frac{\partial B_\varphi}{\partial z}\right). \tag{22}$$

Equating Equations (17) and (22), we obtain the second order differential equation of B_φ as

$$\frac{\partial^2 B_\varphi}{\partial z^2} - \left(\frac{4\pi e^2 N_{oe}}{\omega_o^2 m_e}\frac{\partial B_\varphi}{\partial z}\right) - \left(k_\dagger^2 - \frac{\omega^2}{c^2} + \frac{\omega_p^2}{c^2}\right)B_\varphi + \left(\frac{4\pi e^2 k_\perp \varphi \nabla n}{\omega_o c m_e}\right) = 0. \tag{23}$$

The laser frequency "ω_o" is in the order of 10^{14} Hz, which fabricates the constant term of $\frac{\partial B_\varphi}{\partial z}$ almost negligible. In the context of a plasma slab featuring sharp boundaries, there is a surface current responsible for causing a discontinuity in the magnetic field. It is worth noting that this discontinuity in the magnetic field, while present, is typically of lesser

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

significance when contrasted with the discontinuity in $\frac{\partial B_\varphi}{\partial z}$, which often leads to its omission in calculations. Equation (23) reduces to the form

$$\frac{\partial^2 B_\varphi}{\partial z^2} - \left(k_\dagger^2 - \frac{\omega^2}{c^2}\varepsilon_z\right)B_\varphi + \left(\frac{4\pi e^2 k_\perp \varphi(\nabla \mu)}{\omega_o c m_e}\right) = 0. \tag{24}$$

On solving Equation (23) in vacuum ($z < -L, z > L$) and in plasma ($-L < z < L$), and by applying the continuity conditions for the radial and azimuthal components of electric and magnetic fields respectively, at the interface between the plasma and vacuum, we can determine the characteristics of the low-frequency electromagnetic field across the entire columns of space. The wave entering from the left side of the slab and running away toward the right side of the slab provides us with the equations mentioned as

$$B_\varphi = C_3 e^{(\lambda_o\{z+L\})}, \quad (z < -L), \tag{25}$$

$$B_\varphi = C_4 e^{(-\lambda_o\{z-L\})}, \quad (z > L), \tag{26}$$

$$B_\varphi = C_1 e^{\lambda z} + C_2 e^{-\lambda z} + \frac{4\pi e^2 k_\perp \varphi}{K^2 \omega_o c m_e}\nabla n, \quad (-L < z < L), \tag{27}$$

where $K^2 = \left(k_\dagger^2 - \frac{\omega^2}{c^2}\varepsilon_z\right)$, and C_1, C_2, C_3, and C_4 are constants.

In order to calculate the boundary conditions for Equation (14), that is, "E_r" for the plasma-vacuum interface, we use

$$\nabla \times \nabla \times \vec{E} = -\frac{i\omega_o}{c}\frac{\partial B_\varphi}{\partial z}. \tag{28}$$

Substituting Equation (28) in Equation (14), we get E_r in terms of B_φ as

$$E_r = \frac{-ic}{\omega_o \varepsilon_z}\left(\frac{\omega_p^2 k_\perp \varphi}{\omega c} + \frac{\partial B_\varphi}{\partial z}\right). \tag{29}$$

Applying the continuity conditions and boundary conditions at the above equation, we receive the equations in plasma and in vacuum of the form:

$$E_r = \frac{-c\lambda_o}{\omega} C_3 e^{(-\lambda_o\{z+L\})} \text{for } (z < -L), \tag{30}$$

$$E_r = \frac{c\lambda_o}{\omega} C_4 e^{(\lambda_o\{z-L\})} \text{for } (z > L), \tag{31}$$

$$E_r = \frac{-ic}{\omega\varepsilon_z}\left[\frac{\omega_p^2 k_\perp \varphi}{\omega c} + C_1\lambda e^{\lambda L} - C_2\lambda e^{-\lambda L} - \frac{16\pi^2 e^4 k_\perp}{\omega_o c^3 m_e^2 \lambda^4}(\nabla n)^2\right] \text{for } (-L < z < L). \tag{32}$$

Equating Equations (30) and (32) as well as Equations (25) and (27) at the left side of the plasma-vacuum interface, and Equations (31) and (32) as well as Equations (26) and (27) at the right side of the plasma-vacuum interface, we obtain the values of constants as

$$C_3 = C_1 e^{-\lambda L} + C_2 e^{\lambda L} + \frac{4\pi e^2 k_\perp \varphi \nabla n}{\omega_o c m_e}, \tag{33}$$

$$C_4 = C_1 e^{\lambda L} + C_2 e^{-\lambda L} + \frac{4\pi e^2 k_\perp \varphi \nabla n}{\omega_o c m_e}. \tag{34}$$

Equating the constants C_3 and C_4 for both radial component of electric field and azimuthal component of magnetic field, we obtain relations

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$C_1(\lambda + i\lambda_o\varepsilon_z)e^{-\lambda L} - C_2(\lambda - i\lambda_o\varepsilon_z)e^{\lambda L} + \varphi\left[\frac{\omega_p^2 k_\perp}{\omega c} + i\lambda_o c - \frac{16\pi^2 e^4 k_\perp}{\omega_o c^3 m_e^2 \lambda^4}(\nabla n)^2\right], \tag{35}$$

$$C_1(\lambda - i\lambda_o\varepsilon_z)e^{\lambda L} - C_2(\lambda + i\lambda_o\varepsilon_z)e^{-\lambda L} + \varphi\left[\frac{\omega_p^2 k_\perp}{\omega c} - i\lambda_o c - \frac{16\pi^2 e^4 k_\perp}{\omega_o c^3 m_e^2 \lambda^4}(\nabla n)^2\right]. \tag{36}$$

Adding and subtracting Equations (35) and (36) simultaneously, we obtain

$$\begin{aligned} &C_1\left[(\lambda + i\lambda_o\varepsilon_z)e^{-\lambda L} + (\lambda - i\lambda_o\varepsilon_z)e^{\lambda L}\right] - C_2\left[(\lambda - i\lambda_o\varepsilon_z)e^{\lambda L} + (\lambda + i\lambda_o\varepsilon_z)e^{-\lambda L}\right] \\ &\quad + 2\varphi\left[\frac{\omega_p^2 k_\perp}{\omega c} + \frac{16\pi^2 e^4 k_\perp}{\omega_o c^3 m_e^2 \lambda^4}(\nabla n)^2\right] = 0, \end{aligned} \tag{37}$$

$$\begin{aligned} &C_1\left[(\lambda + i\lambda_o\varepsilon_z)e^{-\lambda L} - (\lambda - i\lambda_o\varepsilon_z)e^{\lambda L}\right] - C_2\left[(\lambda - i\lambda_o\varepsilon_z)e^{\lambda L} - (\lambda + i\lambda_o\varepsilon_z)e^{-\lambda L}\right]. \\ &\quad + 2\varphi(i\lambda_o\varepsilon_z) = 0 \end{aligned} \tag{38}$$

Within these mathematical expressions, the terms represented along with "C_1" and "$C2$" are typically recognized as "T_s" and "T_a" respectively, given as

$$C_1 = \left[\frac{\omega_p^2 k_\perp}{\omega c} - \frac{16\pi^2 e^4 k_\perp}{\omega_o c^3 m_e^2 \lambda^4}(\nabla n)^2\right]\left(\frac{-\varphi}{T_s}\right) + i\lambda_o\varepsilon_z\left(\frac{\varphi}{T_a}\right), \tag{39}$$

$$C_2 = \left[\frac{\omega_p^2 k_\perp}{\omega c} - \frac{16\pi^2 e^4 k_\perp}{\omega_o c^3 m_e^2 \lambda^4}(\nabla n)^2\right]\left(\frac{-\varphi}{T_s}\right) - i\lambda_o\varepsilon_z\left(\frac{\varphi}{T_a}\right), \tag{40}$$

$$T_s = (\lambda + i\lambda_o\varepsilon_z)e^{(-\lambda L)} + (\lambda - i\lambda_o\varepsilon_z)e^{(\lambda L)}, \tag{41}$$

$$T_a = (\lambda - i\lambda_o\varepsilon_z)e^{(\lambda L)} - (\lambda + i\lambda_o\varepsilon_z)e^{(-\lambda L)}. \tag{42}$$

The terms "T_s" and "T_a" correspond to the excitation of the plasma slab's electromagnetic eigenmodes (symmetric and antisymmetric), as they would appear in denominators of electric field (Equations 30–32) and magnetic field expressions (Equations 25–27). The dispersion relationship governing the symmetric electromagnetic mode within a plasma slab can be formulated when a particular condition is met which is determined by $T_s = 0$. This condition yields the following dispersion relationship:

$$\frac{\lambda}{\varepsilon}\tanh(\lambda L) + i\lambda_o = 0. \tag{43}$$

Similarly, the antisymmetric electromagnetic mode within a plasma slab can be formulated when $T_a = 0$. This yields the following dispersion relation for antisymmetric mode given by:

$$\frac{\lambda}{\varepsilon}\coth(\lambda L) + i\lambda_o = 0. \tag{44}$$

Equation (44) describes the dispersion relation for a symmetric mode that travels within a plane characterized by uniform plasma density, which exponentially diminishes as one moves away from the plasma boundary. This equation is a crucial mathematical expression used to understand how this specific mode behaves under these unique conditions. It helps us grasp the intricate relationship between the mode's propagation properties, the consistency of plasma density, and the rapid decay of plasma characteristics as we move away from its edge. Analyzing this equation can provide essential insights into the wavelength, frequency, and phase velocity of the mode in question, contributing to the understanding of wave-plasma interactions and their relevance in various scientific and technological contexts.

3 | RESULTS AND DISCUSSION

We estimate terahertz radiation field for typical parameters of laser-plasma experiments. A femtosecond laser pulse with intensity 10^{12}–10^{16} W/cm^2 is incident on a 80-μm thick plasma slab having electron number density of 10^{17}–10^{19} cm^{-3}.

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The laser beam is incident on the left boundary of plasma slab at $z = -40\,\mu m$. The laser propagates through the plasma slab and tunnels out of the right boundary of the slab at $z = 40\,\mu m$. We observe the THz field pattern and intensity for all the laser profiles under study and for all the three regions of laser propagation, that is, $z < -L$, $z > L$, and $-L < z < L$, where $L = 40\,\mu m$, and L is also varied to observe the THz field variation with plasma slab thickness. The gradient of the laser field (spatial variation) plays a crucial role in generating the electric radial field. When the laser intensity varies spatially within the plasma slab, it creates spatial gradients in the ponderomotive potential. These gradients lead to forces on charged particles within the plasma, causing them to move and generate the radial electric field. Essentially, the particles are pushed by the ponderomotive force associated with the gradient of the laser field.

Figure 1 shows the variation of normalized electric field as a function of radial distance r, for different spatially polarized profiles viz. Gaussian, super-Gaussian, cosh-hyperbolic, sinusoidal, triangular, and rectangular-triangular waveforms. Due to varying profile envelopes, the amplitude of incident laser varies with radial distance. The field amplitude of laser is also different for different profiles and is maximum for cosh-hyperbolic waveform. All other profiles have the same maximum amplitude value of $10^6\,V\,cm^{-1}$, while cosh-hyperbolic laser beam has two maximum points of $1.3 \times 10^6\,V\,cm^{-1}$ each on either side of $r = 0$ symmetrically. The amplitude at these two points depends on the value of decentered parameter "b" of laser beam profile. The laser propagates through the plasma and imparts an oscillatory velocity to the plasma electrons, which exerts a ponderomotive force on them, pushing them forward until it behaves as the space charge force. Since the laser beam is radially polarized, therefore the electrons experience a radial ponderomotive force and drive a nonlinear current in azimuthal direction.

We have plotted the variation of ponderomotive potential "ϕ" with radial distance "r" in Figure 2. The ponderomotive potential, influenced by laser intensity, mirrors the laser's spatial pattern as it is a quiver energy that describes how charged particles respond to the time-averaged laser field. When the laser intensity varies, the ponderomotive potential also varies accordingly. As the ponderomotive potential oscillates, it imparts kinetic energy to charged particles in the plasma, causing them to oscillate and contribute to the electric radial field. This potential, experienced by charged particles in a plasma, is proportional to the laser's power per unit area. Its gradient induces electric field excitation both within and outside the plasma slab accordingly to Equations (30–32). This phenomenon plays a pivotal role in shaping charged particle trajectories in laser-plasma interactions. The understanding of ponderomotive potential and it is associated dynamics is crucial in various applications, such as laser-driven fusion research and particle acceleration.

The nonlinear current excites a magnetic field in azimuthal direction. The variation of the self-generated magnetic field along the direction of laser propagation and the gradient of ponderomotive potential in radial direction, together contribute to the radial component of electric field. The axial component of the excited electric field is also related to the azimuthal magnetic field and is directly proportional to it is radial gradient. In Figure 3, we have plotted the normalized radial electric field within the plasma slab as a function of radial variable "r" in 2-D for all the laser

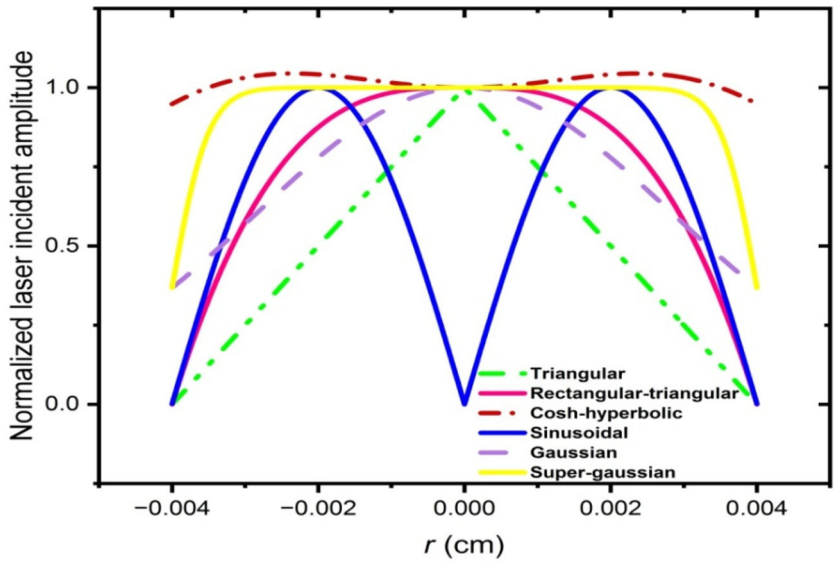


FIGURE 1 Normalized incident laser field amplitude as a function of "r" for rectangular-triangular, triangular, cosh-hyperbolic, sinusoidal, Gaussian, and super-Gaussian profiles.

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

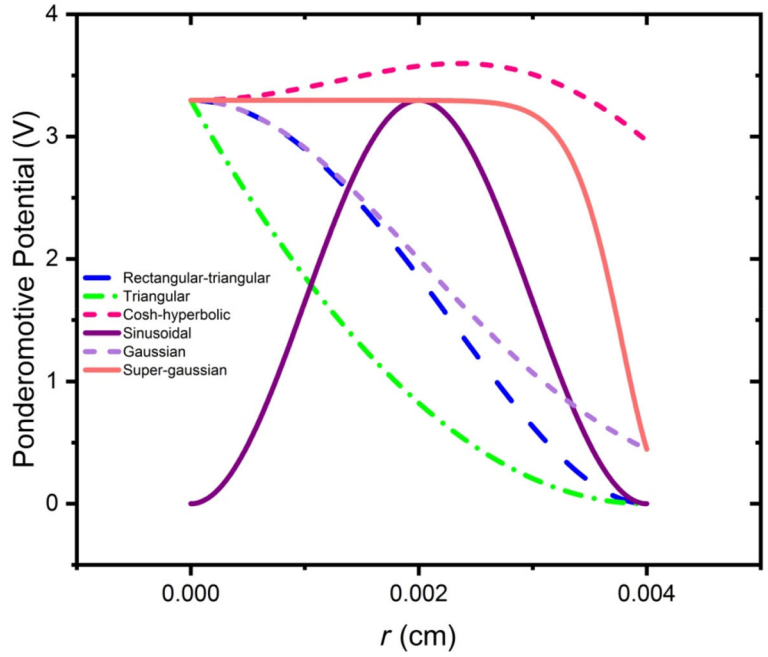


FIGURE 2 Ponderomotive potential as a function of "r" for different laser profiles.

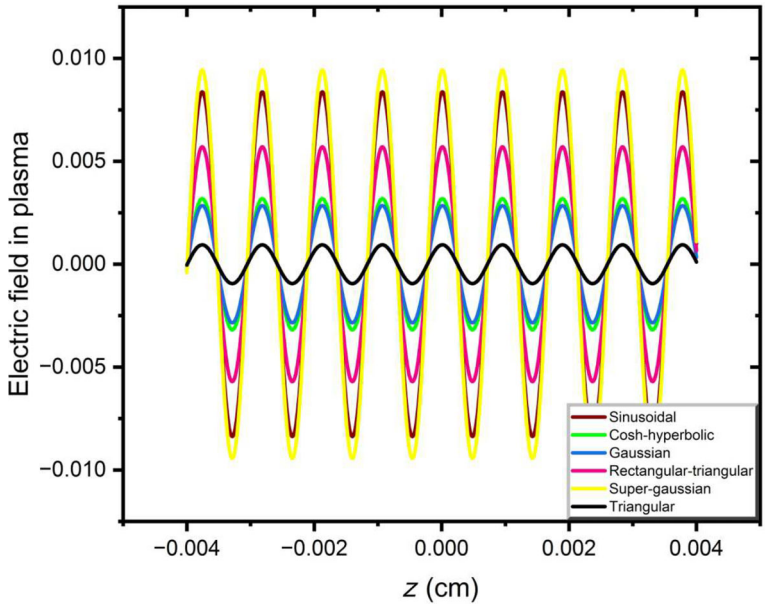


FIGURE 3 Normalized radial electric field "E_r" in plasma slab for different laser profiles.

profiles under consideration. It can be seen that normalized electric field is maximum for Super-Gaussian laser profile beam, which is then followed by rectangular-triangular, sinusoidal, cosh-hyperbolic, and triangular profiles. This variation maybe understood on comparing Figures 2 and 3, which indicate that the radial electric field is maximum where the slope of the ponderomotive potential versus r graph is maximum.

To get an insight of the variation of radial electric field with both spatial coordinates "r" and "z", we have simulated the 3D contour plots of the excited field within the plasma slab for limits $-L < z < L$ and outside the slab, on both sides in vacuum for all the laser beam profiles. Figure 4 shows the contour plots of normalized electric field within the plasma slab as a function of spatial variables "r" and "z," for sinusoidal, triangular, rectangular-triangular, cosh-hyperbolic ($b = 1.2$), cosh-hyperbolic ($b = 2.0$), Gaussian, and super-Gaussian laser pulse profiles, respectively. The gradient of the

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

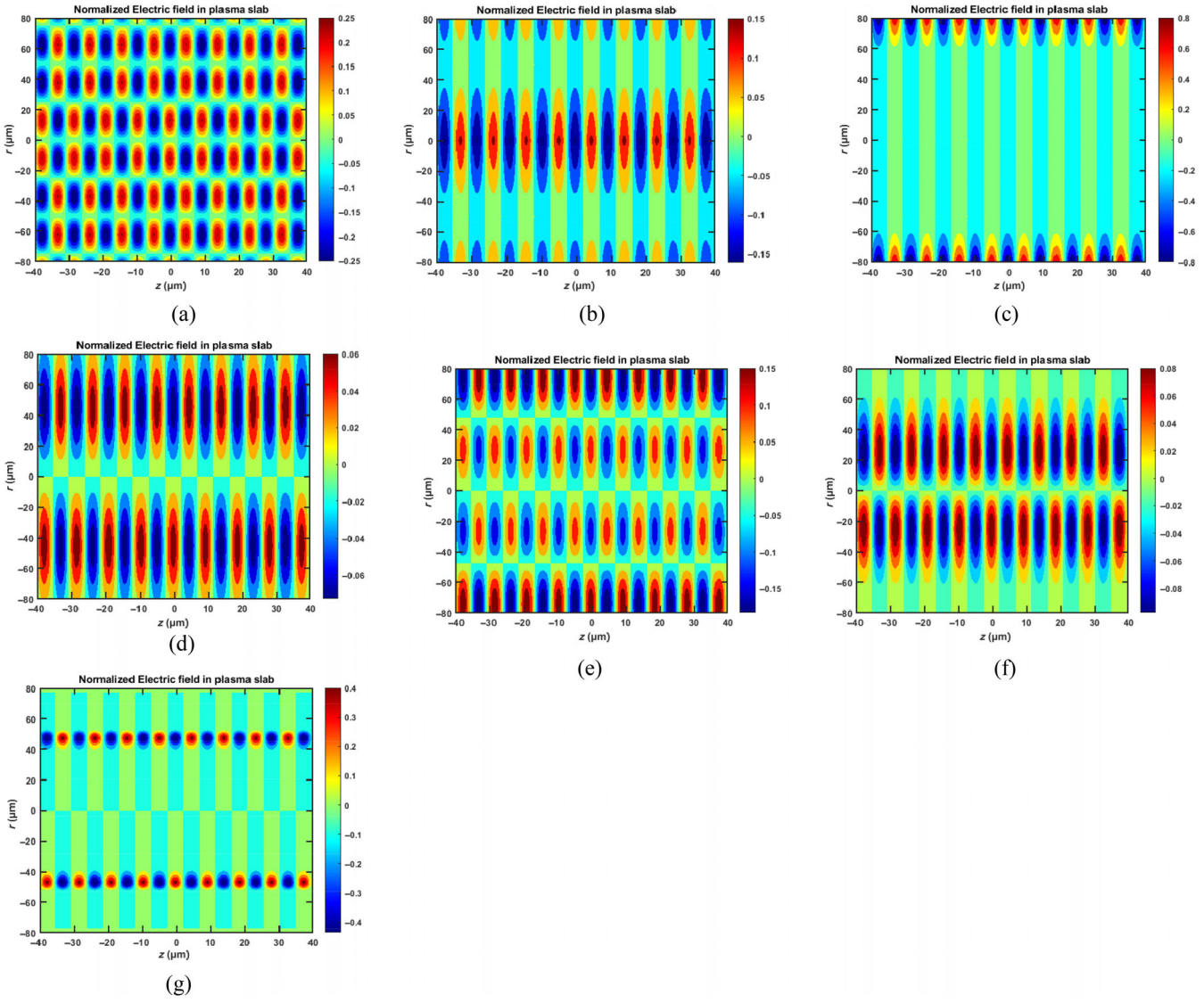


FIGURE 4 Normalized radial electric field "E_r/E_{ol}" in plasma slab for (a) sinusoidal, (b) triangular, (c) rectangular-triangular, (d) cosh-hyperbolic ($b = 1.2$), (e) cosh-hyperbolic ($b = 2.0$), (f) Gaussian, and (g) super-Gaussian laser profiles. Other parameters are $n_e = 1.3 \times 10^{19}$ cm^{-3}, $L = 40$ μm, $E_{ol} = 10^8$ V/cm, $\omega_o = 2 \times 10^{15}$ rad/s, $\omega = 2.07 \times 10^{14}$rad/s, $\theta = 10°$, and $b_w = 50$ μm.

laser field within the plasma slab leads to spatial variations in the ponderomotive potential, which in turn, results in forces on charged particles. These forces cause the particles to oscillate and generate the electric radial field. The normalized radial electric field (E_r/E_{ol}) in a plasma slab for different laser profiles varies in shape and magnitude, with Gaussian profiles producing a Gaussian E_r distribution, rectangular-triangular profiles yielding a more uniform E_r within the beam's central region, and other profiles exhibiting unique E_r patterns, all subject to the specific laser-plasma interaction physics involved. The rectangular-triangular laser ($s = 3$) generates maximum peak electric field of 0.8×10^8 V/cm due to it is flat-top pattern, succeeded by super-Gaussian laser ($n = 4$) with peak value of electric field as 0.4×10^8 V/cm, followed by sinusoidal laser with 0.25×10^8 V/cm, cosh-hyperbolic laser ($b = 2$) with 0.15×10^8 V/cm, triangular laser with 0.15×10^8 V/cm, Gaussian laser with 0.08×10^8 V/cm, and the least peak electric field for cosh-hyperbolic laser ($b = 1.2$) with 0.06×10^8 V/cm. Such a variation is justified as the rectangular-triangular and super-Gaussian lasers have steep radial gradient of ponderomotive potential as compared with other profiles. Hematizadeh et al.[63] and Gupta et al.[64] have also reported that top-hat laser beams and flattened Gaussian laser beams produce a stronger ponderomotive force and an enhanced THz field strength. Super-Gaussian and rectangular triangular beams can generate higher radial electric fields in laser-plasma interactions as they have a flatter central region compared with other beam profiles, which means that they can deliver more laser energy to a smaller, well-confined spatial area.

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

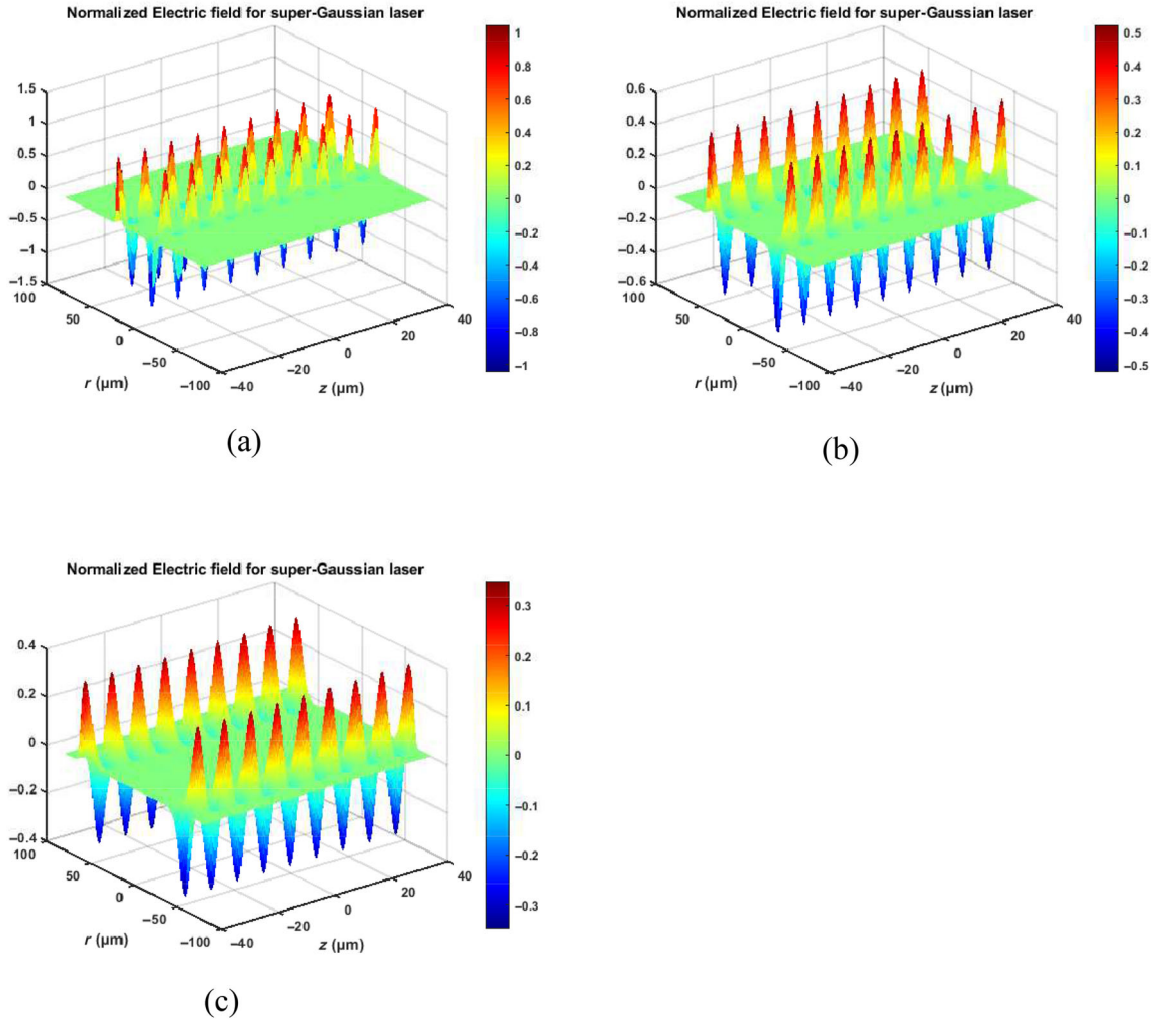


FIGURE 5 Variation of normalized electric field in plasma slab for super-Gaussian laser profile beam with "r" and "z" for different laser pulse width: (a) $b_w = 20\,\mu m$, (b) $b_w = 40\,\mu m$, and (c) $b_w = 60\,\mu m$.

Figure 5 illustrates the three-dimensional variation of normalized THz electric field with spatial coordinates, for different values of laser beam width $b_w = 20$, 40, and 60 μm, with rest of the parameters same as in Figure 4. The laser profile chosen here is super-Gaussian. One sees that the excited radial THz field is more focused and higher for lower values of beam width. A two-fold enhanced THz field is obtained when beam width is halved. Figure 5a depicts a highly focused beam for $b_w = 20\,\mu m$ as compared with a less focused beam for $b_w = 60\,\mu m$ in Figure 5c. It follows from the fact that for a laser pulse of small beam width, the characteristic wavenumber "$k_\perp$" which is inversely proportional to beam width b_w, is fairly large, and therefore, the waves propagate primarily in the radial direction. Thus, a decrease in the beam width enhances the efficiency of excitation in radial direction. Frolov[65] also reported that mainly surface waves are excited, when a laser pulse with tight focus falls on a plasma half-space boundary and therefore the waves predominantly propagate in radial direction.

Figure 6 shows the behavior of normalized THz electric field with frequency for varying plasma slab thickness, that is, $L = 25$, 35, 45, and 55 μm with laser frequency taken as $w_o = 2 \times 10^{14}$ rad/s and $n_e = 1.3 \times 10^{17}\ cm^{-3}$. The excited radial electric fields are emitted at THz frequencies and increase with decreasing slab thickness. Though the intensity of THz field decreases, but the field spectrum increases with the increase in value of L, and the frequency of maxima also increases lying within the range 15 to 20 THz. A 2D plot of THz electric field as a function of frequency, for different laser beam widths is shown in Figure 7 with the same parameters as taken for Figure 6. Reduction in beam width increases the intensity of emitted THz radiation but the frequency of maxima remains almost same for all values of beam width, which is comparable with the plasma frequency. As the plasma pulse width decreases, the area under the curve increases, indicating a wider spectrum with more intense excited electric field. The results are in conjunction with the 3D plots shown in Figure 5.

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

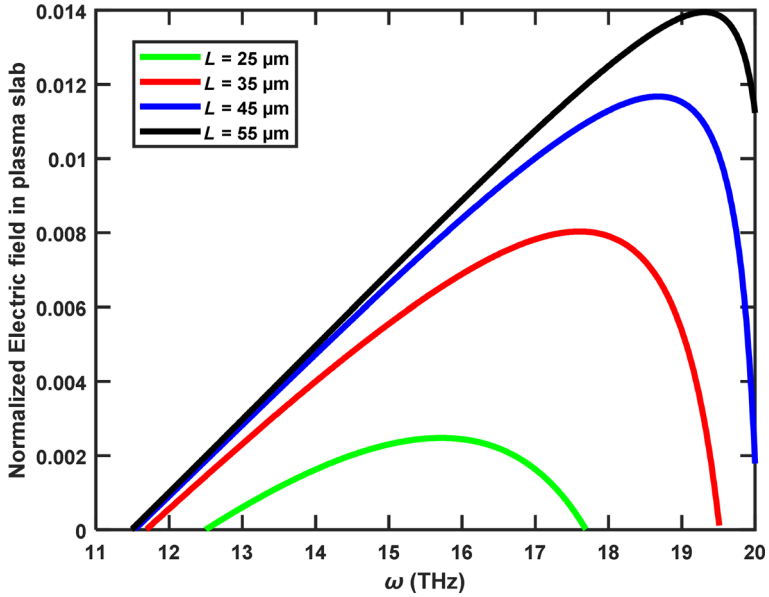


FIGURE 6 Normalized electric field in plasma slab for sinusoidal laser waveform as a function of THz frequency for different slab thickness $L = 25$, 35, 45, and 55 μm.

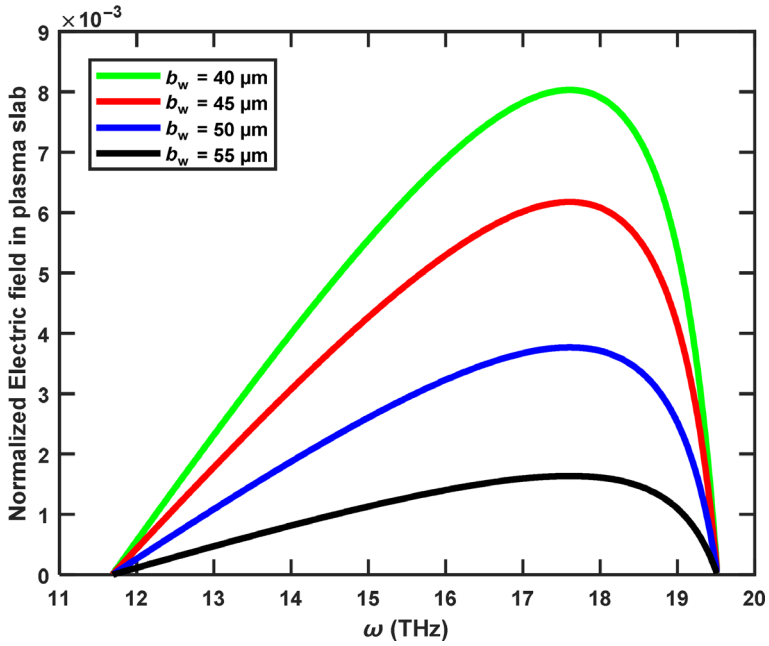


FIGURE 7 Normalized electric field as a function of THz frequency for varied laser beam widths for a sinusoidal waveform laser.

As a result of complex plasma currents, the plasma slab borders behave like antennas and produce electromagnetic radiation in a conical form. Therefore, we have also studied the angular distribution of THz field inside the plasma slab and in vacuum. Figure 8 shows normalized THz field as a function of frequency for different viewing angles. Maximum THz field is observed at a perspective angle of 45° to the laser propagation direction. For angles less than 45°, an increase in the view angle increases the intensity as well as spectrum of THz field and the frequency of maximum intensity field remains close to the plasma frequency at about 17 THz. For angles more than 45°, an increase in the view angle, decreases the emitted THz field intensity but the THz spectrum expands with view angle. These observations are specifically for Gaussian laser pulse profile. For other laser profiles also, a similar pattern of variation is observed. It is shown from

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

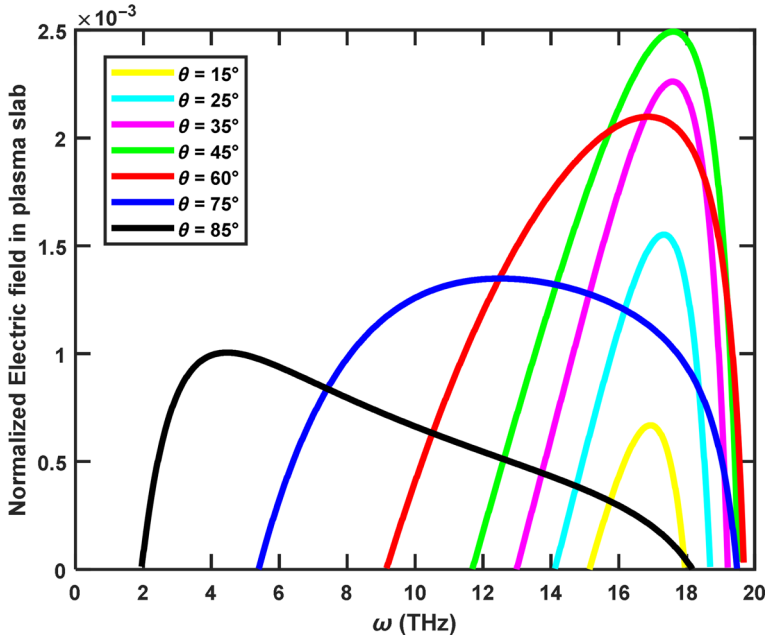


FIGURE 8 Normalized THz field response across various viewing angles as a function of frequency for sinusoidal wave.

Figure 6 that as the plasma slab thickness decreases, the spectrum of radiation also decreases. When the slab thickness is comparable with the laser beam width, the intensity of excited field is the highest as shown in Figure 7. Similar results have been concluded by Frolov et al.[65] for terahertz radiation generation from a low-density plasma slab with the similar plasma parameters as considered here for plotting Figures 6–8.

The relationship between the incident laser amplitude and the generated magnetic field is nonlinear. This means that small changes in laser intensity can lead to disproportionately larger changes in the generated magnetic field. As the intensity of the incident laser pulse increases, the electrons in the plasma are subjected to stronger forces due to the electric field of the laser. The laser pulse's electric field can accelerate free electrons within the plasma, leading to the formation of electron currents. These electron currents are associated with the motion of charged particles and, according to Ampere's law, can generate magnetic fields. Linearly polarized lasers generate magnetic fields perpendicular to the laser's polarization direction, while circularly polarized lasers can produce magnetic fields in different directions. We also report magnetic field in azimuthal direction produced by radially polarized laser pulses. In our study, we create contour plots illustrating the strength of the magnetic field for various laser pulse profiles. Figure 9 depicts the magnetic field pattern in plasma slab for sinusoidal, Gaussian, super-Gaussian, cosh-hyperbolic, triangular, and rectangular-triangular laser pulse envelopes. Through the analysis of these patterns, we observed that the magnetic field reaches highest intensity for super-Gaussian pulse, while it exhibits its lowest strength for triangular pulse waveform. Similar pattern of field strength was obtained from radial electric field contours also for individual laser profiles. The physical mechanism responsible for the generation of azimuthal component of the magnetic field is governed by the ponderomotive forces. The magnetic field azimuthal component is proportional to the square of laser amplitude. The spatial variation of the magnetic field strongly depends on the laser pulse shape and density of plasma.

Figure 10 exhibits the relationship between incident laser amplitude and the self-generated magnetic field. Researchers[66] have identified scaling laws that describe the relationship between laser intensity and the resulting magnetic field. These laws may provide insights into the behavior of the system. Our measured magnetic field values closely align with the values predicted by scaling laws formulated by other researchers. When measuring or observing the generated magnetic field, the angle at which the measurement is taken relative to the laser's propagation direction can affect the observed field strength. The Doppler shift due to the motion of electrons in the plasma can affect the observed frequency and strength of the magnetic field when measured at different angles. The variation of magnetic field with slab thickness is influenced by intensity of the laser, the electron number density of the plasma, and the duration of the laser pulse. Figure 11 depicts the variation of self-generated magnetic field with plasma slab thickness for three different perspective angles. When a laser pulse interacts with a plasma, it ionizes the plasma creating charged particles (ions and free electrons). A thicker plasma slab implies higher concentration of ions and electrons, which can enhance the generation of

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

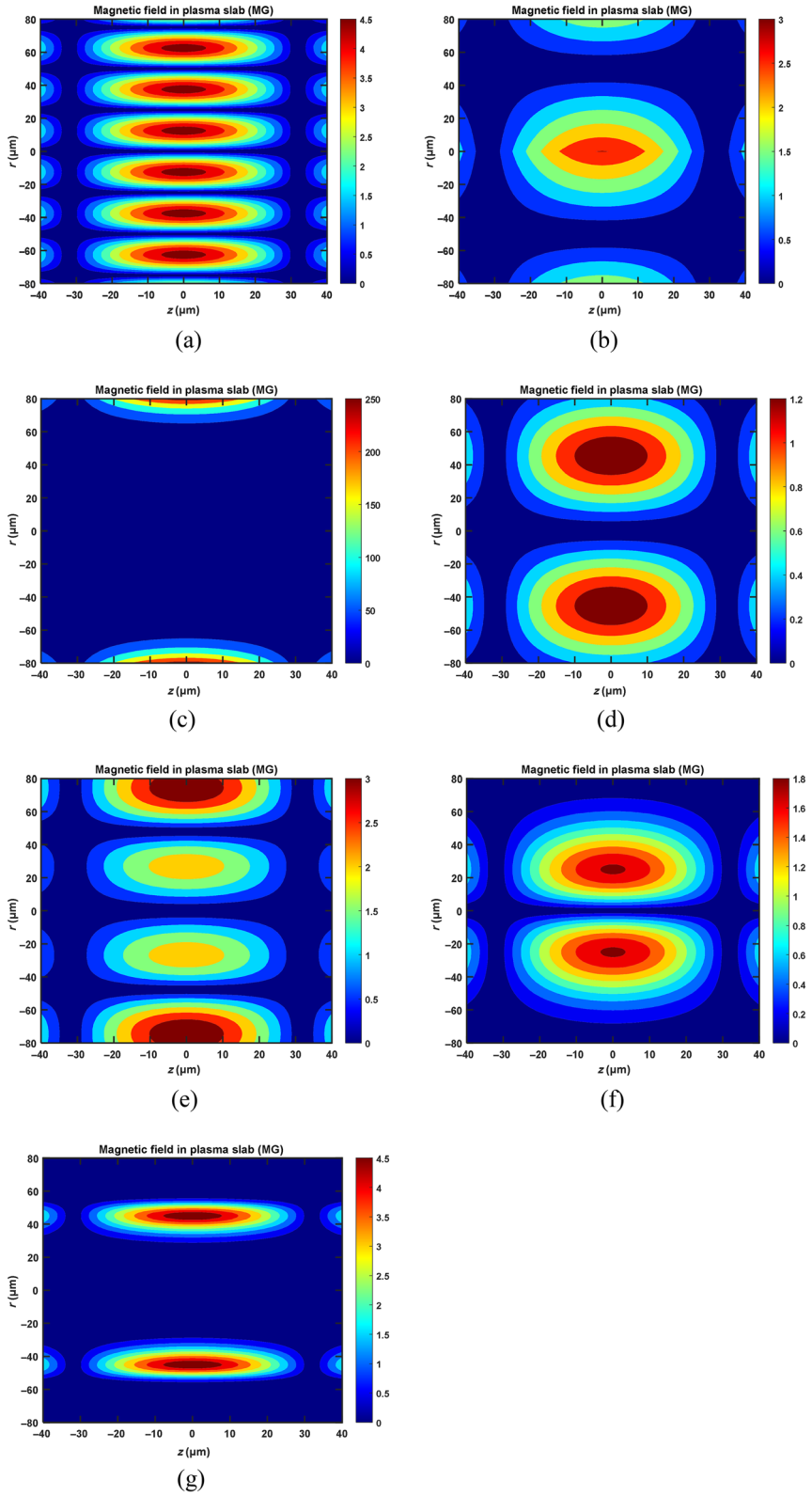


FIGURE 9 Self-generated magnetic field "B_Φ" in plasma slab for (a) sinusoidal, (b) triangular, (c) rectangular-triangular, (d) cosh-hyperbolic ($b = 1.2$), (e) cosh-hyperbolic ($b = 2.0$), (f) Gaussian, and (g) super-Gaussian laser profiles. Other parameters are $n_e = 1.3 \times 10^{19}$ cm^{-3}, $L = 40$ μm, $E_{ol} = 10^8$ V/cm, $\omega_o = 2 \times 10^{15}$ rad/s, $\omega = 2.07 \times 10^{14}$ rad/s, $\theta = 10°$, and $b_w = 50$ μm.

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

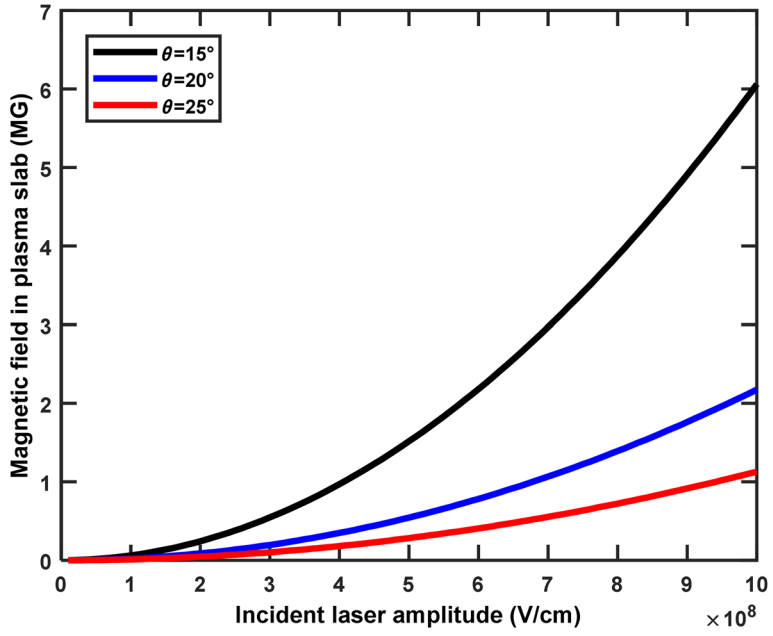


FIGURE 10 Self-generated magnetic field "B_Φ" in plasma slab as a function of incident super-Gaussian laser amplitude E_{ol} (V/cm) for different viewing angles θ.

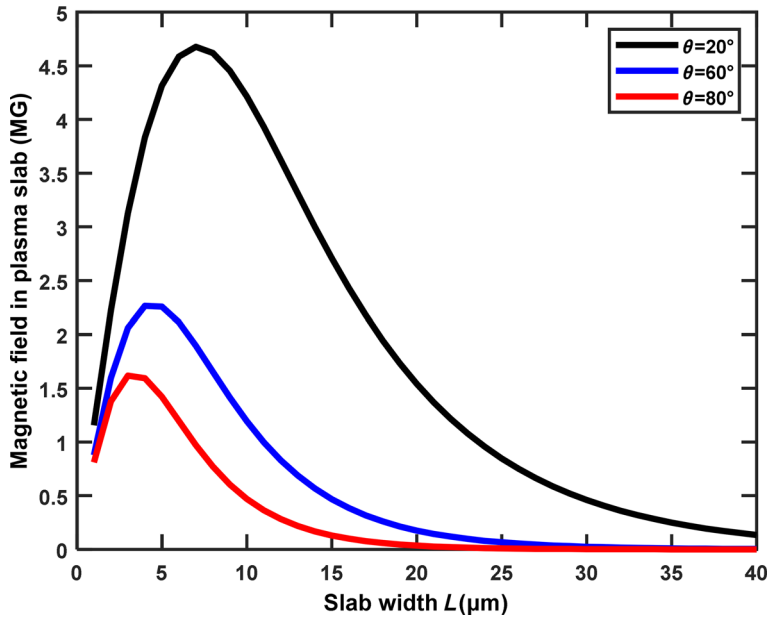


FIGURE 11 Self-generated magnetic field "B_Φ" in plasma slab as a function of plasma slab width "L" on interaction with Gaussian laser pulse for different perspective angles θ.

magnetic fields. However, as the plasma slab becomes very thick, saturation effects are seen. Beyond a certain thickness, the magnetic field strength starts declining. The value of saturation magnetic field and the corresponding slab thickness depends on the detailed properties of the laser pulse profile and the angle at which azimuthal magnetic field is viewed from the center of the slab. Figure 11 shows that as the view angle decreases, the magnetic field strength decreases for thick slabs($L > 15\,\mu m$) while it increases for thin slabs($L < 5\,\mu m$). At a 20° view angle, the maximum observed magnetic field strength is 4.7 MG.

To conduct a comparative analysis of our findings in the context of existing research, we employed the code of Gaussian laser system, utilizing parameters as specified by previous experimental and theoretical studies. At an intensity

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of 10^{17} W cm^{-2}, Santos et al.[5] were able to generate a magnetic field of 10 MG, when the laser interacted with solid state targets. Lamac et al.[66] reported a generation of azimuthal magnetic field in GG range in their experimental work for high laser amplitudes; a reduction in laser intensity will generate magnetic fields obtained from our mathematical analysis, also justifying the numerical scaling laws. For plasma with electron density $\sim 10^{18}$ cm^{-3}, Gorbunov and Ramazashvili[67] using perturbation theory reported a maximum magnetic field of 350 kG, which is comparable with the field values observed in our analysis. Tatarakis et al.[27] experimentally measured peak magnetic fields in the range 340–460 MG with a laser intensity greater than 10^{19} W cm^{-2}, for higher harmonics. We observed magnetic fields in the range 1–5 MG with laser intensities of 10^{16} W cm^{-2}, which are justifiable as our intensity of laser is comparatively less and we report the results for fundamental harmonic frequency. Park et al.[68] numerically computed a magnetic field of 0.6 GG, but for higher plasma densities and a highly intense laser beam, which is also comparable with our results.

4 | CONCLUSION

We propose an analytical model to investigate the excited radial electric field and self-generated magnetic field in a homogeneous plasma slab during interaction with a radially polarized laser beam with varied pulse profiles such as sinusoidal, Gaussian, cosh-hyperbolic, super-Gaussian, triangular, and rectangular-triangular wave forms. 2D and 3D contour plots are employed to scrutinize the intensity of the electromagnetic fields through simulation. Through a comparison between the observed results and the provided experimental and theoretical framework, the concept is effectively and convincingly exemplified. Super-Gaussian and rectangular triangular beams can generate higher radial electric fields in laser-plasma interactions due to their unique intensity profiles. They have a flatter central region compared with other beam profiles, which means that they can deliver more laser energy to a smaller, well-confined spatial area. The enhanced intensity in the central region of a super-Gaussian beam can trigger stronger nonlinear effects when interacting with the plasma. These nonlinear effects can result in the generation of higher harmonics and more complex plasma responses, including stronger radial electric fields. It is important to note that the specific behavior of varied profile beams in laser-plasma interactions can vary depending on the beam parameters, the plasma conditions, and the desired outcome of the interaction. The choice of beam profile, whether Gaussian, super-Gaussian, sinusoidal, or another profile, depends on the specific experimental or theoretical objectives in laser-plasma physics. The electric vector of a radially polarized laser beam is parallel to radial coordinate and can produce strong longitudinal electric fields if focused.[69] Such electric fields have small laser spot size and useful for optical trapping of particles. Rectangular-triangular, super-Gaussian and sinusoidal lasers exhibit a significantly steeper radial gradient of ponderomotive potential in comparison with other laser profiles, thus exciting intense radial electric fields and generating high order azimuthal magnetic fields in plasma. The cone angle over which the field is emitted expands as the thickness of the slab increases. Decreasing the beam width results in an augmentation of the intensity of emitted THz radiation, while the frequency of the maxima remains nearly constant across various beam width values, remaining comparable with the plasma frequency. The underlying physical mechanism responsible for generating the azimuthal component of the magnetic field is predominantly governed by ponderomotive forces and the symmetric eigenmode in plasma slab. The antisymmetric eigenmode resulted in very low values of excited fields, making the symmetric mode as dominant one. In our analysis of field patterns, we noted that the electromagnetic fields attain peak intensity with the rectangular-triangular and super-Gaussian pulse waveforms, while they demonstrate lowest strength when using the triangular pulse waveform. The saturation magnetic field value, as well as the associated slab thickness, are contingent on the specific characteristics of the laser pulse profile and the angle from which the azimuthal magnetic field is observed relative to the center of the slab. We have employed a nonrelativistic model which is valid as long as the dimensionless laser amplitude $\left(a_o = \frac{eE_{ol}}{m_e c\omega}\right)$ is less than 1. The numerical results presented in this study align closely with the findings obtained in other theoretical and simulation studies, and find numerous applications.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

REFERENCES

[1] J. X. Liu, Y. J. Liu, H. W. Yang, *Int. J. Aeronaut. Space Sci.* **2023**, *24*, 618.
[2] J. Ji, Y. Ma, N. Guo, *Optik* **2018**, *165*, 240.
[3] M. Yamada, H. Ji, S. Hsu, T. Carter, R. Kulsrud, N. Bretz, F. Jobes, Y. Ono, F. Perkins, *Phys. Plasmas* **1997**, *4*, 1936.
[4] D. Biskamp, Magnetic Reconnection in Plasmas, Cambridge University Press, Cambridge **2000**.

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[5] J. J. Santos, M. B. Grandvaux, M. Ehret, A. V. Arifiev, D. Batani, F. N. Beg, A. Calisti, S. Ferri, R. Florido, P. F. Colleoni, S. Fujioka, M. A. Gigosos, L. Giuffrida, L. Gremillet, J. J. Honrubia, S. Kojima, P. Korneev, K. F. F. Law, J. R. Marques, A. Morace, C. Mosse, O. Peyrusse, S. Rose, M. Roth, S. Sakata, G. Schaumann, F. S. Vidal, V. T. Tikhonchuk, T. Toncian, N. Woolsey, Z. Zhang, *Phys. Plasmas* **2018**, *25*, 056705.
[6] D. J. Strozzi, M. Tabak, D. J. Larson, L. Divol, A. J. Kemp, C. Bellei, M. M. Marinak, M. H. Key, *Phys. Plasmas* **2012**, *19*, 072711.
[7] M. G. Haines, K. Krushelnick, R. J. Kingham, A. R. Bell, A. E. Dangor, M. S. Wei, A. Gopal, M. Tatakaris, F. N. Beg, P. A. Norreys, R. G. Evans, E. L. Clark, R. J. Clark, *20th IAEA fusion energy conf. proc.* **2005**, 3451.
[8] D. Biskamp, *Nucl. Fusion* **1973**, *13*, 719.
[9] M. Herrmann, M. Tani, M. Watanabe, K. Sakai, *IEE Proc. Optoelectronics* **2002**, *149*, 116.
[10] A. D'Arco, M. Di Fabrizio, V. Dolci, M. Petrarca, S. Lupi, *Condens. Matter* **2020**, *5*, 25.
[11] B. Born, M. Havenith, *J. Infrared Millim. Terahertz Waves* **2009**, *30*, 1245.
[12] A. Dinovitser, D. G. Valchev, D. Abbott, *R. Soc. Open Sci.* **2017**, *4*, 170725.
[13] M. Heyden, E. Brundermann, U. Heugen, G. Niehues, D. M. Leitner, M. Havenith, *J. Am. Chem. Soc.* **2008**, *130*, 5773.
[14] I. F. Akyildiz, J. Miquel, C. Han, *Phys. Commun.* **2014**, *12*, 16.
[15] J. Zhang, S. Li, W. Le, *iScience* **2021**, *3*, 103548.
[16] J. Zhang, S. Li, W. Le, *I Science* **2021**, *24*, 103548.
[17] A. L. Herrera-May, L. A. Aguilera-Cortés, P. J. García-Ramírez, E. Manjarrez, *Sensors* **2009**, *9*, 7785.
[18] K. Gopal, D. N. Gupta, H. Suk, *IEEE Trans. Plasma Sci.* **2021**, *41*, 1152.
[19] T. Tajima, X. Q. Yan, T. Ebisuzaki, *Rev. Mod. Plasma Phys.* **2020**, *4*, 7.
[20] C. W. Siders, S. P. Le Blanc, D. Fisher, M. C. Downer, A. Babine, A. Stepanov, T. Tajima, Ultrafast Phenomena X, Vol. *62*, *Springer*, Berlin, Heidelberg **1996**.
[21] R. Lundin, A. Guglielmi, *Space Sci. Rev.* **2006**, *127*, 1.
[22] C. Meng, W. Chen, X. Wang, Z. Lü, Y. Huang, J. Liu, J. M. Yuan, D. W. Zhang, X. Sun, H. Z. Wu, Y. Y. Tu, P. Song, Z. Y. Zhou, C. S. Meng, *Appl. Phys. Lett.* **2016**, *109*, 131105.
[23] L. Yan-Mei, L. Ya-Ning, L. Zhang, J. Zhi-Hong, Z. Song-Feng, W. Guo-Li, *Front. Phys.* **2023**, *11*, 1.
[24] Z. Xiao-B, M. Huang, R. R. Tang, Z. Ai-Xia, X. Ju-Kui, *AIP Adv.* **2022**, *12*, 055003.
[25] V. V. Ivanov, A. V. Maximov, R. Betti, L. S. Leal, J. D. Moody, K. J. Swanson, N. A. Huerta, *Matter Radiat. Extrem.* **2021**, *6*, 4.
[26] Q. Z. Luo, N. D. Angelo, R. L. Merlino, *Phys. Plasmas* **1999**, *6*, 3455.
[27] M. Tatarakis, I. Watts, F. N. Beg, E. L. Clark, A. E. Dangor, M. G. Haines, P. A. Norreys, U. Wagner, M. S. Wei, M. Zepf, K. Krushelnick, *Nature* **2002**, *415*, 280.
[28] G. J. Pert, *J. Plasma Phys.* **1977**, *18*, 227.
[29] J. A. Stamper, K. Papadopulos, R. N. Sudan, S. O. Dean, E. A. McLean, J. M. Dawson, *Phys. Rev. Lett.* **1971**, *26*, 1012.
[30] A. S. Sandhu, A. K. Dharmadhikari, P. P. Rajeev, S. Sengupta, A. Das, P. K. Kaw, G. R. Kumar, *Phys. Rev. Lett.* **2002**, *89*, 225002.
[31] O. M. Gradov, L. Stenflo, *Phys. Lett.* **1983**, *95*, 233.
[32] V. Thakur, N. Kant, S. Vij, *Phys. Scr.* **2020**, *95*, 045602.
[33] S. Vij, N. Kant, V. Thakur, *Plasmonics* **2019**, *14*, 1051.
[34] A. Mehta, N. Kant, *Proc. SPIE OPTO* **2019**, *10917*, 1.
[35] D. J. Cook, R. M. Hochstrasser, *Optics Lett.* **2000**, *25*, 1210.
[36] A. Proulx, A. Talebpour, S. Petit, S. L. Chin, *Optics Commun.* **2000**, *174*, 305.
[37] X. Xie, J. Dai, X. C. Zhang, *Phys. Rev. Lett.* **2006**, *196*, 075005.
[38] A. Houard, Y. Liu, B. Prade, V. T. Tikhonchuk, A. Mysyrowicz, *Phys. Rev. Lett.* **2008**, *100*, 255006.
[39] Z. Zhang, Y. Chen, M. Chen, Z. Zhang, J. Yu, Z. Sheng, J. Zhang, *Phys. Rev. Lett.* **2016**, *117*, 243901.
[40] M. Clerici, M. Peccianti, B. E. Schmidt, L. Caspani, M. Shalabay, M. Giguere, A. Lotti, A. Couairon, F. Legare, T. Ozaki, D. Faccio, R. Morandotti, *Phys. Rev. Lett.* **2013**, *110*, 253901.
[41] L. L. Zhang, W. M. Wang, T. Wu, R. Zhang, S. J. Zhang, C. L. Zhang, Y. Zhang, Z. M. Sheng, X. C. Zhang, *Phys. Rev. Lett.* **2017**, *119*, 235001.
[42] V. A. Andreeva, O. G. Kosareva, N. A. Panov, D. E. Shipilo, P. M. Solyankin, M. N. Esaulkov, P. G. A. Martinez, A. P. Shkurinov, V. A. Makarov, L. Berge, S. L. Chin, *Phys. Rev. Lett.* **2016**, *116*, 063902.
[43] V. B. Gildenburg, N. V. Vvedenskii, *Phys. Rev. Lett.* **2007**, *98*, 245002.
[44] H. C. Wu, J. M. T. Vehn, Z. M. Sheng, *New J. Phys.* **2008**, *10*, 043001.
[45] M. Chen, A. Pukhov, X. Y. Peng, O. Willi, *Phys. Rev. E* **2008**, *78*, 046406.
[46] W. M. Wang, Z. M. Sheng, H. C. Wu, M. Chen, C. Li, J. Zhang, K. Mima, *Opt. Express* **2008**, *16*, 16999.
[47] K. Y. Kim, J. H. Glownia, A. J. Taylor, G. Rodriguez, *Opt. Express* **2007**, *13*, 4577.
[48] A. K. Malik, H. K. Malik, *IEEE J. Quantum Electron* **2013**, *49*, 232.
[49] F. Bakhtiari, S. Golmohammady, M. Yousefi, F. D. Kashani, B. Ghafary, *Laser Part. Beams* **2015**, *33*, 463.
[50] C. F. Phelan, J. F. Donegan, J. G. Lunney, *Opt. Express* **2011**, *19*, 21793.
[51] M. Navarro-Cia, J. Wu, H. Liu, O. Mitrofanov, *Sci. Rep.* **2016**, *6*, 38926.
[52] T. Wakayama, O. G. Rodriguez-Herrera, J. Scott Tyo, Y. Otani, M. Yonemora, T. Yoshizawa, *Opt. Express* **2014**, *22*, 3306.
[53] P. Ma, P. Zhou, Y. Ma, X. Wang, R. Su, Z. Liu, *Opt. Lett.* **2012**, *37*, 2658.
[54] J. Mansell, T. Rutherford, W. Tulloch, M. Olapinski, M. Fejer, R. Byer, *Thursday Morning CLEO*. **2000**, 406.
[55] C. Y. Huang, K. C. Chang, S. C. Chu, *Materials* **2019**, *12*, 2226.

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[56] J. Kim, P. Jeon, H. Lee, D. Y. Kim, *Opt. Lasers Eng.* **2024**, *176*, 108053.
[57] L. L. Zhang, W. M. Wang, T. Wu, S. J. Feng, K. Kang, C. L. Zhang, Y. Zhang, Y. T. Li, Z. M. Sheng, X. C. Zhang, *Phys. Rev. Appl.* **2019**, *12*, 014005.
[58] Y. Tan, H. Zhao, W. M. Wang, R. Zhang, Y. J. Zhao, C. L. Zhang, X. C. Zhang, L. L. Zhang, *Phys. Rev. Lett.* **2022**, *128*, 093902.
[59] H. K. Malik, T. Punia, D. Sharma, *Electronics* **2021**, *10*, 3134.
[60] F. Bakhtiari, M. Esmaeilzadeh, B. Ghafary, *Phys. Plasmas* **2017**, *24*, 073112.
[61] M. Kumar, V. K. Tripathi, *IEEE J. Quantum Electron.* **2012**, *48*, 1031.
[62] X. P. Hu, P. Xu, S. N. Zhu, *Photon. Res.* **2013**, *1*, 171.
[63] A. Hematizadeh, M. Jazayeri, B. Ghafary, *Contrib. Plasma Phys.* **2018**, *1*, 1.
[64] D. N. Gupta, A. Jain, V. V. Kulagin, M. S. Hur, H. Suk, *Appl. Phys. B* **2022**, *128*, 49.
[65] A. A. Frolov, *Plasma Phys. Rep.* **2007**, *33*, 179.
[66] M. Lamač, U. Chaulagain, J. Nejdl, S. V. Bulanov, *Sci. Rep.* **2023**, *13*, 1701.
[67] L. M. Gorbunov, R. R. Ramazashvili, *J. Exp. Theor. Phys.* **1998**, *87*, 461.
[68] J. Park, S. S. Bulanov, J. Bin, S. Steinke, J. L. Vay, G. R. Geddes, C. B. Schroeder, W. P. Leemans, T. Schenkel, E. Esarey, *Phys. Plasmas* **2019**, *26*, 103108.
[69] D. N. Gupta, H. Suk, *J. Appl. Phys.* **2007**, *101*, 043109.

How to cite this article: H. Sagar, S. C. Sharma, *Contrib. Plasma Phys*. **2024**, e202400020. https://doi.org/10.1002/ctpp.202400020

15213986, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ctpp.202400020 by Delhi Technological University, Wiley Online Library on [13/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ACADEMIA MATERIALS SCIENCE

Review Article
Published: 2024-05-02
https://doi.org/10.20935/AcadMatSci6184

Role of polymers in encapsulation and shape-stabilization of phase change materials (PCMs) for thermal energy storage applications: an overview

Surya Tanwar[1], Raminder Kaur[1,*]

Academic Editor(s): Islam Md Rizwanul Fattah, Raul D.S.G. Campilho

Abstract

Phase change materials (PCMs) are suitable for thermal energy storage (TES) owing to their increased heat storage ability, narrow temperature difference between storage and retrieval of latent heat energy, and readiness in a variety of phase change transition temperatures. Numerous research attempts have been made to develop a broad array of PCMs based on inorganic, organic, and polymeric compounds for a diverse area of applications. To improve the performance efficiency and proper handling of PCMs, they are often encapsulated or confined within a shape-stabilized supporting material of an inorganic or organic origin. This review emphasizes the current research status of the polymer-based PCMs developed using the various PCMs–polymer combinations available, material characteristics, encapsulation, and shape-stabilization methods and provides detailed information on the contribution of polymers toward the enhancement of PCM end-product properties, in an avant-garde technology related to TES.

Keywords: *phase change materials, microencapsulation, shape-stabilized, thermal energy storage*

Citation: Tanwar S, Kaur R. Role of polymers in encapsulation and shape-stabilization of phase change materials (PCMs) for thermal energy storage applications: an overview. *Academia Materials Science* 2024;1. https://doi.org/10.20935/AcadMatSci6184

1. Introduction

The transition in the manufacturing practices, advancements in technology, and the population boom have led to an enormous increase in energy demand in the past few centuries. Over the last half-century, researchers around the world are seeking for new fuel alternatives due to less availability of fossil fuels in the earth's reservoirs and a sudden surge in fuel prices. Moreover, awareness among the people toward the environmental and climatic changes has also inspired researchers and investigators to explore different renewable energy sources. Proposed theories that advocate solar radiation as most appreciably a viable energy source become unrealistic, as solar energy is extremely intermittent in nature leading to discrepancy between the availability and utilization time. The solution to all these energy problems is to stock the energy in the form that can be converted as and when required. The energy storage ensures the performance and steadfastness of energy systems and reduces the discrepancy between the supply and demand.

Thermal energy storage (TES) systems based on phase change materials (PCMs) have gained much attention in the research community over the last few decades. PCMs have latent heat storage properties in which the latent heat energy transfer takes place when a material transforms from the solid to a liquid state or vice versa. Most of the available literature related to PCMs for TES applications reveals the various PCMs contained inside a matrix or supporting material of inorganic or organic nature. Based on a thorough survey made, it has been explored that polymers play a vital role as matrix materials in producing PCM products for TES applications owing to their improved physical and thermal properties, good volume and leakage control, along with a remarkable dimensional stability, especially required for solid–liquid PCMs. This inclusive review presents state-of-the-art PCMs for TES applications, different polymer shells, or polymer supporting matrices, which are used for their preparation. It gives evidence about their enhanced performance and safety, and provides information about the efforts of different researchers in this area.

2. Thermal energy storage (TES)

TES is a temporary storage of high- or low-temperature energy for later use and can be retrieved by heating (melting), cooling (solidifying), or vaporizing a material with energy, which is reflected in the form of heat, when the process is reversed. TES methods (**Figure 1**) can be classified into three broad categories, namely (1) sensible heat storage, i.e., due to temperature change or change in internal energy of the material; (2) latent heat storage, based on the heat capture or release during the phase change of the storage material; and (3) thermo-chemical heat storage, relying on the heat energy derived from the chemical reactions of materials, which are used for storage purposes [2].

[1]Department of Applied Chemistry, Delhi Technological University, Shahbad Daulatpur, Delhi, 11042, India.
*email: raminderkaur@dce.ac.in

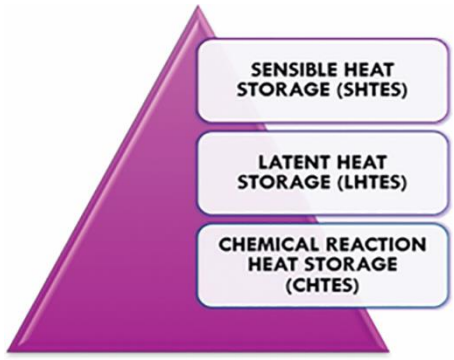


Figure 1• Different types of thermal energy storage (TES) methods. SHTES, sensible heat thermal energy storage; LHTES, latent heat thermal energy storage, and CHTES, chemical reaction heat thermal energy storage.

2.1. Sensible heat thermal energy storage (SHTES) systems

Sensible heat thermal energy storage (SHTES) encompasses the storage of thermal energy by elevating or diminishing the temperature of the storage medium, i.e., solid or liquid. The energy loading potential is corresponding to the amount of the storage material, its specific heat, density, thermal conductivity, and temperature variations during the initial and final stages [3]. The sensible heat stored in any material can be calculated using equation 1:

$$Q = \int_{T_i}^{T_f} mC_p dT = mC_{ap}(T_f - T_i) \quad (1)$$

where Q is the quantity of heat stored (J), T_i is the initial temperature (°C), T_f is the final temperature (°C), m is the mass of the heat storage medium (kg), C_p is the specific heat (J/kg K), and C_{ap} is the average specific heat between T_i and T_f (J/kg K). SHTES has less energy density and entails a huge capacity for the storage that leads to a relatively bigger unit size. These are ideal for small applications, where economics is a leading concern. The conventionally used SHTES materials, since olden times, are sand, rocks, concrete, bricks, water, oils, metals, molten salts, etc. Among these, water is the best SHTES material, as it is cost-free, abundant, and has high specific heat capacity.

2.2. Chemical reaction heat thermal energy storage (CHTES) systems

Chemical reaction heat thermal energy storage (CHTES) works on the principle of the energy absorbed and released in a reversible chemical reaction, by the breaking and re-forming of molecular bonds. These are more beneficial as compared to SHTES due to their high energy densities, ambient temperature storage, and the possibility of heat pumping and long-distance transport. The amount of energy stored is in correspondence with the amount of reacting materials, the extent of conversion, and the endothermic heat of reaction [2]. The amount of energy stored in a thermo-chemical energy storage system can be calculated using equation 2:

$$Q = a_r m \Delta h_r \quad (2)$$

where Q is the quantity of the heat stored (J), m is the mass of the heat storage medium (kg), Δh_r is the heat of reaction, and a_r is the fraction reacted.

2.3. Latent heat thermal energy storage (LHTES) systems

In latent heat thermal energy storage (LHTES) systems, the energy storage process takes place by absorbing and retrieving the thermal energy in the form of latent heat, when the storage material undergoes a change in physical state or phase, i.e., from solid to liquid, liquid to gas, or vice versa. The energy storage of LHTES systems can be described by equation 3:

$$Q = \int_{T_i}^{T_m} mC_p dT = ma_m \Delta H_m + \int_{T_i}^{T_m} mC_p dT \quad (3)$$

where a_m is the fraction of the material melted, ΔH_m is the heat of melting per unit mass (J/kg), and m is the mass of the heat storage medium (kg). The PCMs constitute a significant category of the LHTES system.

3. Phase change materials (PCMs)

Latent heat storage with the aid of PCMs is the highly effective way of thermal energy management owing to their higher energy storage density, with a narrower and almost constant temperature ranges [1]. PCMs work upon the basic principle of phase transformation, when they encounter a temperature swing. The TES occurs in the form of latent heat when the storage material shows phase alteration, typically, among solid–liquid, solid–solid, liquid–gas, or solid–gas, etc. For example, when a solid–liquid PCM is exposed to the heating process, its temperature rises constantly. As soon as the temperature of the storage material approaches the melting point, it begins to melt completely and uniformly. As the melting process is endothermic, it absorbs and stores the thermal energy available in its surroundings in the form of latent heat, in a liquid state. Furthermore, when the cycle is reversed, during the cooling process, with the progressive fall in temperature, the PCM liberates an equal amount of stored energy in the surroundings and transforms back into a solidified state. During the entire heat–cool cycle, the temperature remains constant during the phase change [4]. **Figure 2** represents a schematic outline of the mechanism of the phase change process in a solid–liquid PCM.

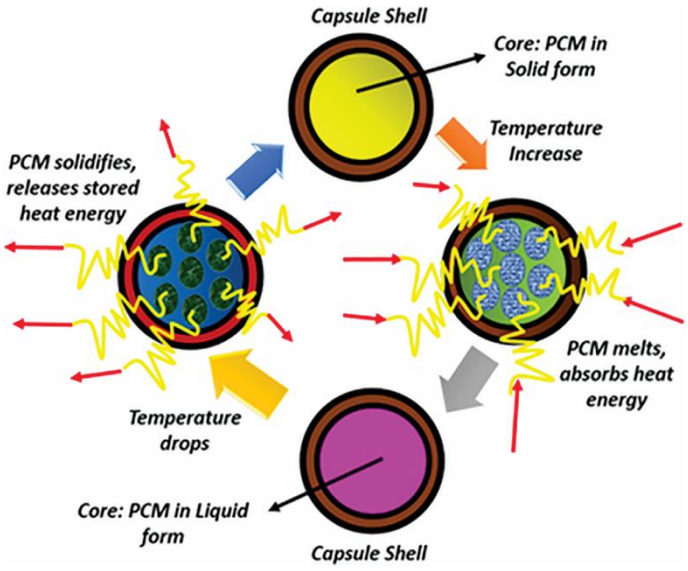


Figure 2 • A schematic outline of the mechanism of the phase change process in a solid–liquid PCM.

The energy storage based on PCMs possesses certain merits such as (1) storing a large quantity of latent heat per unit volume, i.e., high energy storage densities; (2) phase change at a constant temperature, i.e., uniform latent heat transfer; and (3) a broad spectrum of PCMs is available with a wide range of operating temperatures, which enable to develop products with tunable temperatures as per the required application [5]. With the information compiled from different studies, the main characteristic requirements for an ideal PCM for TES applications are as follows:

- ✓ Phase change temperature in the domain of functioning temperature of the heating or cooling of the product to attain optimum performance.
- ✓ High latent heat energy per unit volume to store more thermal energy.
- ✓ High thermal conductivity to support higher heat transfer rates and facilitate rapid charging and discharging.
- ✓ Uniform melting at a constant temperature.
- ✓ Endure repeatable phase transformation cycles for a long period.
- ✓ Negligible or low supercooling.
- ✓ High nucleation rate to promote a faster solidification process and the occurrence of reversible cycles at nearly the same temperature.
- ✓ Experience only a small volume change during the phase transitions to maintain the geometry of the container in-housing them.
- ✓ Low vapor pressure to overcome the risk of rupture of encapsulating the material due to high-pressure build-up situations.
- ✓ High density, which facilitates higher energy storage capacity.
- ✓ Chemical stability to ensure long life and non-reactivity with the encapsulating matrix to avoid the formation of undesirable products and leakage from containers.
- ✓ Non-corrosive to avoid the deterioration of containers.
- ✓ Non-flammable to avoid the risk to building materials and textiles in which it is encapsulated.
- ✓ Non-hazardous and non-poisonous to prevent poisonous emissions during a fire or in case of rupture of the encapsulation.
- ✓ Easy availability and economical.

The different requisite characteristic properties of PCMs for TES applications have been summarized in **Table 1**.

Table 1 • Basic requirements of PCMs for TES applications [6]

Thermal properties	Kinetic properties	Physical properties	Chemical properties	Socio-economical aspect
Suitable phase-transition temperature (melting point) in desired operating temperatures	Low supercooling	Favorable phase equilibrium—no phase segregation	Chemical stability during phase transformations	Abundant
High latent heat of phase change	High nucleation rate	Small volume change	Complete reversible melting/freezing cycles	Economic
Good thermal conductivity in both solid and liquid phases		Low vapor pressure	Compatible with host matrix/encapsulating shell	
Congruent melting pattern		Reproducible phase change cycles	Non-corrosive	
		High density	Non-flammable	
			Non-toxic	

4. Classification of PCMs

PCMs can be classified broadly into three categories based on (1) temperature, (2) phase change, and (3) chemical nature, as shown in **Figure 3**.

4.1. Based on temperature

The temperature range governing the phase transformation of PCMs falls within the three temperature zones: (1) low-temperature PCMs, with phase change temperatures below 15°C, applied in the fields of air conditioning, cold storage, and food industry; (2) mid-temperature PCMs, with phase change temperatures from 15°C to 90°C, applied in the fields of solar power, electronics, smart textiles, biomedical applications, and energy-efficient buildings; (3) high-temperature PCMs, with phase change temperatures beyond 90°C, applied in the fields of aerospace applications, absorptive refrigeration, and various industrial operations [1, 6].

4.2. Based on phase transitions

Phase transition takes place among solid–liquid, solid–solid, liquid–gas, or solid–gas, etc., and the PCMs can be classified accordingly. The characteristics of each transition have been summarized in **Table 2**.

4.3. Based on chemical nature

Based on chemical nature, PCMs can be categorized as inorganic or organic—each category comprising a broad spectrum of commercial and nascent PCMs. The phase change temperatures and the latent heat of fusions of different inorganic (salt hydrates),

organic (paraffins, fatty acid, PEGs, polyalcohols) eutectics, metals, as reported earlier, are given in **Tables 3–10**. In addition to paraffins and fatty acids, a few more hydrocarbons serve as organic PCMs, as included in **Table 6,** along with their phase change temperatures and the latent heat of fusion.

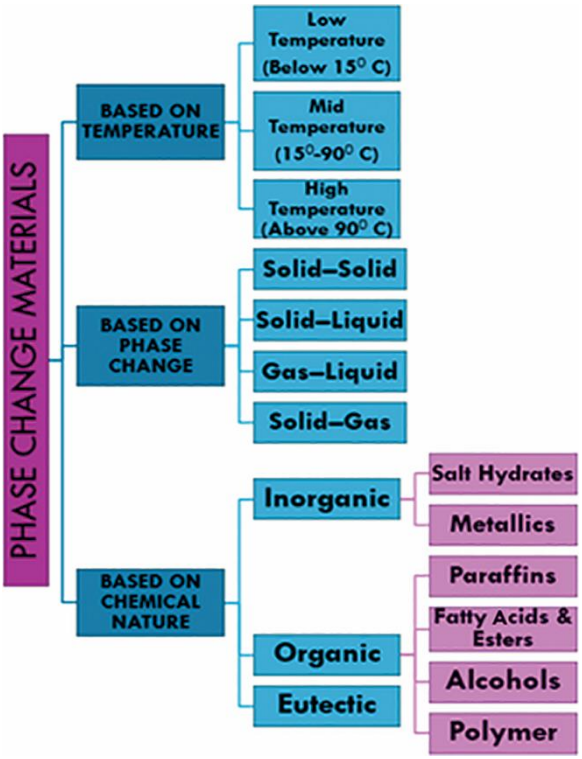


Figure 3 • Classification of PCMs.

Table 2 • Classification of PCMs based on their physical phase transformations

Solid–solid PCMs	Solid–liquid PCMs	Liquid–gas PCMs	Solid–gas PCMs
Involve transformation from crystalline state to amorphous state	Involve transformation from solid state to liquid state	Involve transformation from liquid state to gaseous state	Involve transformation from solid state to gaseous state
Low latent heat energy compared to solid–liquid PCMs	High latent heat energy compared to solid–solid PCMs	Higher latent heat energies	Higher latent heat energies
Small volume change during transitions compared to solid–liquid PCMs	Higher volume change during transitions compared to solid–solid PCMs	Higher volume change during transitions compared to solid–solid PCMs and solid–liquid PCMs	Higher volume change during transitions compared to solid–solid PCMs and solid–liquid PCMs
Need no or lesser encapsulation; leak-proof	Require encapsulation; leak-prone	Require large and bulky containment; leak-prone	Require large and bulky containment; leak-prone
Give better flexibility in design	Practical and economically attractive	Impractical and complex system	Impractical and complex system

Table 3 • Phase change temperatures and latent heat of fusions of salt hydrates

Salt hydrates	Melting point (°C)	Latent heat (kJ/kg)	References
$K_2HPO_4·6H_2O$	14	109	[7]
$Mn(NO_3)_2·6H_2O$	25	148	[8]
$CaCl_2·6H_2O$	28.4	185.6	[9]
$LiNO_3·3H_2O$	30	296	[10]
$Na_2CO_3·10H_2O$	32	267	[11]
$Na_2SO_4·10H_2O$	32.4	241	[12]
$CaBr_2·6H_2O$	34	138	[13]
$Zn(NO_3)_2·6H_2O$	36	136	[14]
$KF·2H_2O$	42	162	[15]
$Na_2SiO_3·5H_2O$	48	168	[16]
$Na_2S_2O_3·5H_2O$	48	210	[17]
$Ca(NO_3)_2·4H_2O$	42	132	[18]
$Co(NO_3)_2·6H_2O$	47	140	[18]
$MnCl_2·4H_2O$	58	151	[19]
$CH_3COONa·3H_2O$	58	270	[20]

Table 4 • Phase change temperatures and latent heat of fusions of paraffins

Alkanes	Formula	Melting point (°C)	Latent heat (kJ/kg)	References
n-Dodecane	$C_{12}H_{26}$	−10	216	[21]
n-Tetradecane	$C_{14}H_{30}$	5–6	226	[22]
n-Pentadecane	$C_{15}H_{32}$	10	207	[23]
n-Hexadecane	$C_{16}H_{34}$	18	236	[24]
n-Heptadecane	$C_{17}H_{36}$	22	215	[25]
n-Octadecane	$C_{18}H_{38}$	28	245	[26]
n-Nonadecane	$C_{19}H_{40}$	32	222	[27]
n-Eicosane	$C_{20}H_{42}$	36–37	247	[28]
n-Heneicosane	$C_{21}H_{44}$	41	215	[29]
n-Tricosane	$C_{23}H_{48}$	47	234	[30]
n-Tetracosane	$C_{24}H_{50}$	51	255	[31]

Table 5 • Phase change temperatures and latent heat of fusions of fatty acids and fatty acid derivatives

Fatty acids and their derivatives	Formula	Melting point (°C)	Latent heat (kJ/kg)	References
Caprylic acid	$CH_3(CH_2)_6COOH$	16	148	[32]
Undecylenic acid	$CH_3(CH_2)_6COOH$	23	141	[33]
Capric acid	$CH_3(CH_2)_8COOH$	32	153	[32]
Lauric acid	$CH_3(CH_2)_{10}COOH$	44	178	[34]
Myristic acid	$CH_3(CH_2)_{12}COOH$	54	205	[35]
Palmitic acid	$CH_3(CH_2)_{14}COOH$	64	212	[36]
Stearic acid	$CH_3(CH_2)_{16}COOH$	70	203	[36]
Propyl palmitate		16–19	186	[37]
Isopropyl palmitate		11	95–100	[38]
Butyl stearate		19	140–200	[39]
Methyl palmitate		29	205	[40]
Methyl stearate		29	169	[41]
Methyl myristate		18	188	[42]
Cetyl alcohol		49	141	[43]
Cetyl palmitate		51	214–220	[44]

Table 6 • Phase change temperatures and latent heat of fusions of some organic PCMs

Organic PCM	Melting point (°C)	Latent heat (kJ/kg)	References
Formic acid	7.8–8.4	247	[45]
Glycerol	18	199	[46]
Caprylone	40	259	[47]
Beeswax	61.8	177	[48]
Acetamide	81	241	[49]
Acetanilide	115	211	[50]
Benzoic acid	124	194	[50]
Hydroquinone	172.4	258	[51]

Table 7 • Phase change temperatures and latent heat of fusions of inorganic and fatty acid eutectics

Inorganic and fatty acid eutectic mixtures	Melting point (°C)	Latent heat (kJ/kg)	References
31% Na_2SO_4 +13% NaCl + 16% KCl + 40% H_2O	4	234	[52]
45% $Ca(NO_3)_2.6H_2O$ + 55% $Zn(NO_3)_2.6H_2O$	25	130	[53]
45–52% $LiNO_3.3H_2O$ + 48–55% Zn $(NO_3)_2.6H_2O$	17.2	220	[54]
47% Ca $(NO_3)_2.4H_2O$ + 53% Mg $(NO_3).6H_2O$	30	136	[55]
66.6% $CaCl_2.6H_2O$ + 33.3% $MgCl_2.6H_2O$	25	127	[53]
50% $CaCl_2$ + 50% $MgCl_2$ + $6H_2O$	25	95	[53]
45% capric acid + 55% lauric acid	17–21	143	[56]
76.5% capric acid+ 23.5% palmitic acid	21.8	171.2	[57]
66% lauric acid + 34% myristic acid	34.2	166.8	[58]
69% lauric acid + 31% palmitic acid	35.2	166.3	[59]
75.5% lauric acid + 24.5% stearic acid	37	182.7	[60]
58% myristic acid + 42% palmitic acid	42.6	169.2	[61]
64% myristic acid + 36% stearic acid	44.1	182.4	[62]
64.2% palmitic acid + 35.8% stearic acid	52.3	181.7	[63]

Table 8 • Phase change temperatures and latent heat of fusions of different molecular weights of PEGs

PEGs	Melting point (°C)	Latent heat (kJ/kg)	References
PEG 400	4.2–5.0	117.6	[64]
PEG 600	12.5	129.1	[65]
PEG 1000	40	168.8	[66]
PEG 1500	46.5	176.3	[67]
PEG 2000	51	181.4	[68]
PEG 4000	59.7	189.7	[69]
PEG 6000	64.8	189	[70]
PEG 10000	66	189.6	[71]
PEG 20000	68.7	187.8	[72]

Table 9 • Phase change temperatures and latent heat of fusions of polyalcohols

Polyalcohol	Melting point (°C)	Latent heat (kJ/kg)	References
Neopentylglycol (NPG)	129	110–120	[73]
2-Amino-2-methyl-1,3-propanediol (AMPL)	100–110	239	[74]
Pentaglycerine (PG)	82	172.5	[75]
Tris(hydroxymethyl)aminomethane (TRIS)	132.4–134.5	285.3–295.6	[76]
Pentaerythritol (PERT)	185–191	289–339.5	[77]

Table 10 • Phase change temperatures and latent heat of fusions of metallic PCMs [7]

Metals	Melting point (°C)	Latent heat (kJ/kg)	References
Gallium	30	80.3	[78]
Bi-Cd-In eutectic	61	25	[6]
Bi-Pd-In eutectic	70	29	[6]
Bi-In eutectic	72	25	[6]

4.4. Encapsulation method

Encapsulation refers to the housing of the material in a container/shell of fixed volume to reach certain objectives [5], such as:

- Prevents undesirable exposure of working with PCMs in an outdoor environment.
- Controls PCM leakage during phase change.
- Increases heat transfer.
- Accommodates volume change during the transition.

The encapsulation method is sub-categorized into (1) macroencapsulation, (2) microencapsulation, and (3) nanoencapsulation.

Macroencapsulation deals with the enclosure of PCMs in a container of any geometry such as pockets, pouches, spheres, shells, thin plates, tubes, and so on. This enables easy installation of PCMs in buildings or act as heat exchangers. The size of macroencapsulated PCMs can reach beyond a few centimeter-level ranges. The features of macroencapsulated PCMs include good leakage prevention and volume control, weathering protection, poor heat transfer, the tendency of solidification at the edges, poor thermal conductivity, etc. [79]. Microencapsulation is a process of embedding solid or liquid PCM materials as core inside a thin layer of polymer or inorganic material as a shell [80], to form a capsule, having a diameter range of 1–1,000 μm [78]. Microencapsulation helps in controlling the leakage of PCMs, volume change during heat cycles, improving heat transfer surface area, adding physical support to the product, easy handling, mitigation of odor or taste, prevention against degradation, and evaporation, among others. Nanoencapsulation is a method of encapsulating PCMs inside a nano-sized entity. Nanoencapsulated PCMs have gained much attention as TES materials in the last two decades due to their high surface area to volume ratio and are much better than microencapsulated PCMs [81].

4.5. Shape-stabilized PCMs

These are another class of TES materials, in which the PCM is dispersed in a supporting matrix to produce dimensionally form-stable composites. These materials possess certain attractive features such as large heat densities, good thermal conductivity, providing shape stability, and leak proofing during phase change, thus eliminating the need for any container for encapsulation and improving the durability over long thermal cycling [80].

5. Tools for the characterization of PCMs

The selection of a suitable PCM for a particular TES application is very crucial. The potential PCM should have a suitable melting temperature, desirable heat of fusion, and thermal conductivity specified for that practical application. Numerous techniques are available for the characterization and measurement of thermal and chemical properties of PCMs. Differential scanning calorimetry (DSC) is used for analyzing the thermal properties of PCMs. In DSC test, the sample and the reference (with known thermal properties) are maintained at the same temperature throughout the measurement process, and by measuring the difference of heat added between the sample and the reference material, thermal properties of the sample can be obtained, such as heat of fusion, heat capacity and melting/solidification temperature, and the effect of multiple thermal cycling on thermal properties of PCMs. The thermal stability of the prepared PCMs can be evaluated using the thermogravimetric analysis (TGA) technique. PCMs are heated over a temperature range from ambient temperature to high elevated temperature profiles (800–900°C) at a prescribed heating rate. Some researchers analyzed the limitations of conventional methods, including conventional calorimetry and DSC, and then put forward a new method called T-history method to determine the melting temperature, degree of supercooling, heat of fusion, specific heat, and thermal conductivity of PCMs. The morphology of the prepared PCM samples can be investigated using a scanning electron microscope (SEM) and particle size distribution (PSD). SEM produces images of a sample by focusing an electron beam on it, and PSD lists the values that define the relative amount of a substance into mixture. PSD sometimes is also called as grain size distribution. Chemical compatibility of the mixture is tested using Fourier transform infrared spectroscopy (FT-IR) technique in which infrared (IR) radiation is passed through the sample. Some of the IR radiation is transmitted through the sample and some absorbed into it. Therefore, to ensure accurate measurement and minimize the uncertainties in the measured values, it is necessary to calibrate the equipment in order to obtain the correct data.

6. Polymer-based PCMs

Polymers are being widely used both for the preparation of encapsulated and shape-stabilized PCMs. The advantages of using polymers for PCM storage are listed in **Figure 4**.

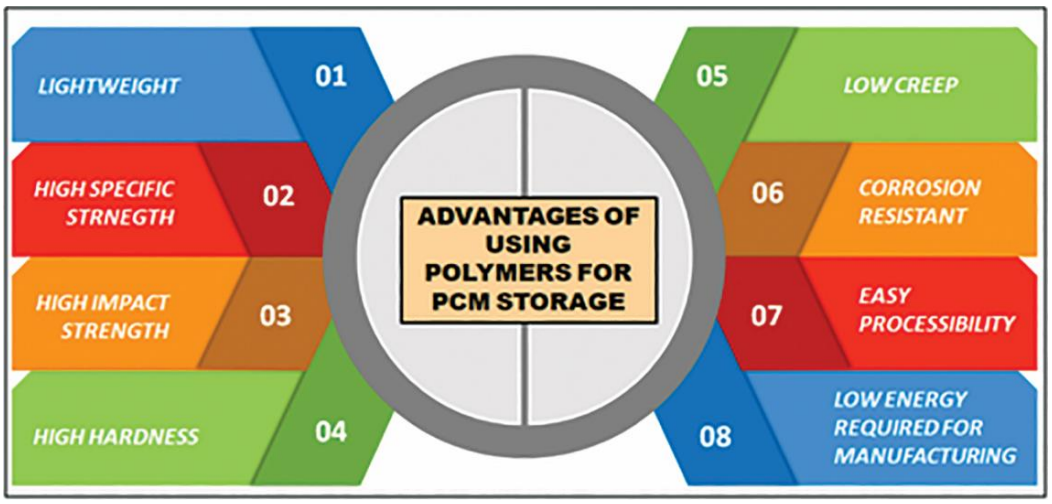


Figure 4 • Benefits of using polymer for PCM storage.

6.1. Polyolefin polymer-based PCMs

Mu et al. [82] produced two different sets of high-density polyethylene (HDPE) shape-stabilized phase change materials (SSPCMs) using twin-screw extruder having low (L-PW, Tm: 18–23°C) and high (H-PW, Tm:56–58°C) melting point paraffin waxes as the TES component. DSC indicated the latent heat enthalpy of SSPCMs with H-PW decreases with increasing H-PW concentration to a higher range due to co-crystallization between H-PW and HDPE. Guo et al. [83] incorporated microencapsulated PCMs (Core: Dodecanol; Shell: MF resin) made by in-situ polymerization into wood–plastic composites (WPCs) based on HDPE matrix to obtain value-added heat storage property. HDPE WPCs containing PCM microcapsules exhibited excellent thermal stability where it undergoes 5% weight loss at 256.9°C. Huang et al. [84] inspected the thermal and structural characteristics of the shape-stabilized composites having organic PCMs such as cetyl alcohol in weight ratio 70–90 wt.% dispersed in HDPE supporting matrix, reinforced by carbon fibers (CF) in varying weight ratio. High dose as much as 5 wt.% CF filled composites showed phase change properties of melting and solidification temperatures at 48.55°C and 41.51°C with accompanied latent heat values of 171.04 and 96.44 kJ/kg, respectively, along with improved leakage control, thermal stability, and thermal conductivity. The addition of expanded graphite (EG) in HDPE containing stearyl alcohol (SAL) as the active PCM, as reported by Tang et al. [85], improved the shape stability of the composites by ensuring good leakage control and an increase in the thermal conductivity from 0.1966 W/(m K) (without EG) to 0.6698 W/(m K) (with 3% EG). The reinforced HDPE composites with SAL mass ratio ~85% had an onset melting temperature of around 57°C with a corresponding latent heat value of 200 kJ/kg.

Halloysite nanotubes (HNTs) were infused with PEG400 and PEG600 to yield form-stable HNT/PCM nano-hybrids mixture, which was melt blended with low-density polyethylene (LDPE) using twin-screw extruder followed by blown film extrusion to finally obtain flexible nanocomposite films with thermal buffering characteristics at low-temperature cold-chain storage and transportation applications. The produced nanocomposite films bestowed reasonable mechanical properties with phase change transition from −17°C to 26°C with associated latent heat of 2.3 J/g. Moreover, nanocomposite films delayed the warming of frozen and chilled samples for 18 and 20 minutes, respectively, relative to neat LDPE films [86]. Premnath et al. [87] used deionized water as the PCM and examined the freezing features of the PCM in LDPE spherical capsule fitted with and without fins. They observed that 50% of the PCM mass has frozen in 21% and 16% of the freezing time with the help of fins when compared to LDPE capsule without fins at −6°C and −9°C bath temperatures.

A shape-stabilized PCM composite composed of HDPE as a packaging material holding paraffin within its macromolecular domain structure containing EG with enhanced thermal conductivity and leakage-proof was prepared. Among all the studied samples in the experiment, the sample made up of HDPE (10 wt.%) and EG (5 wt.%) exhibited the best performance, and its thermal conductivity reached 0.641 W/(mK), which is 2.7 times higher than that of the paraffin. Moreover, the latent heat and thermal cycling test of this sample showed that it can maintain 95% latent heat post 120 thermal cycles, indicating its high thermal storage stability [88].

6.2. Vinyl polymer-based PCMs

Jin et al. [89] fabricated form-stable phase change composite materials (FSPCMs) by compression molding, constituting varying dosage of microencapsulated PCMs (5–25 wt.%) inside polyvinyl chloride (PVC)/wood flour matrix. The mechanical properties were found to be inversely proportional to the PCM microcapsules content, while gradually increasing the weight fraction of microcapsules from 0 to 25 wt.%. The concentration of 10 wt.% of microencapsulated PCMs in PVC/wood flour composites was found to be optimum with good mechanical properties for practical applications. Zhang et al. [90] studied the combined phenomena of the addition of both microencapsulated PCMs as temperature-regulating and TiO_2 nanoparticles as a UV blocking additive to a thin PVC film for solar energy storage applications. The spherical PCM microcapsules of 2–15 µm size range were composed of octadecane as core, well encapsulated inside melamine-formaldehyde resin as a protective shell, synthesized by in-situ polymerization technique. The amount for both PCM microcapsules and TiO_2 nanoparticles in the PVC matrix film was varied according to a weight fraction of 3%, 6%, and 9 wt.%. DSC analysis suggested that PVC films with 6 wt.% octadecane microcapsules showed good heat energy storage density of 8.77 J/g. Further, the incorporation of 6 wt.% TiO_2 nanoparticles to this particular composition of PCM containing films experienced enhanced phase change temperatures and heat storage energy to 9.22 J/g, as compared to other samples. Outcomes of the addition of nano-TiO_2 on the phase change enthalpies and thermal conductivity of the films are shown in **Figure 5**. Temperature-regulating films 6 wt.% micro-PCMs and 6 wt.% TiO_2 nanoparticles displayed excellent thermal reliability where they withstand 100 thermal cycles, with very minute variation phase change temperature and enthalpies as shown in **Figure 6**.

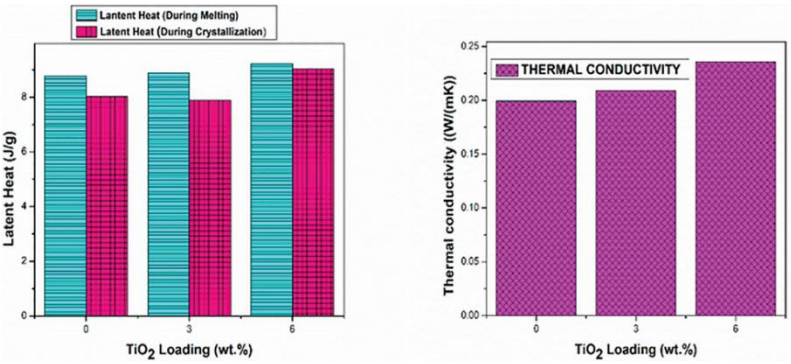


Figure 5 • Influence of nano-TiO_2 addition on enthalpy values and thermal conductivity of the fabricated thermal-regulating films. Data from [90].

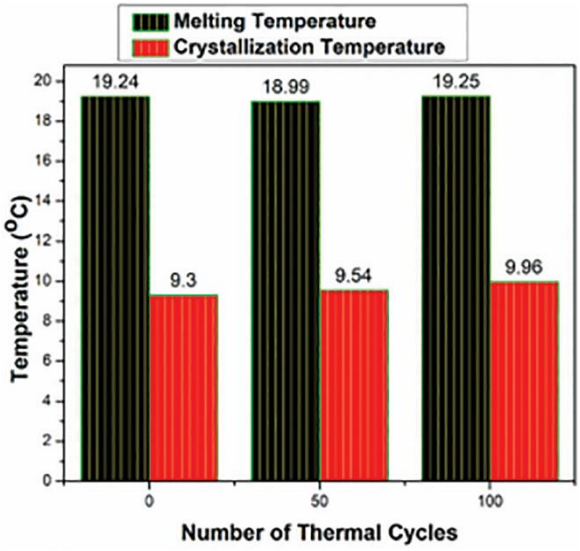


Figure 6 • The thermal cycling stability of the fabricated film incorporated with 6 wt.% micro-PCMs and 6 wt.% nano-TiO_2. Data from [90].

6.3. Styrenic polymer-based PCMs

Giro-Paloma et al. [91] used an emulsion copolymerization technique to obtain a microencapsulated PCM shell made up of poly-(styrene-co-ethyl acrylate) (PScEA) for the encapsulation of paraffins (Capsule A) and palmitic acid (Capsule B) as TES materials. Mean particle sizes of both paraffin and fatty acid encapsulated in copolymer shells were found to be 166 and 265 nm, respectively. AFM data at low and above phase change temperatures represented that Capsule A showed an almost similar force required to collapse the capsules at different temperatures (25°C and 50°C), whereas for fatty acid-loaded Capsule B, the force required to break the capsule at 70°C was ten times than with the amount required at 25°C. Fang et al. [92] worked upon cold TES applications, such as air conditioning, and developed a latent functionally thermal fluid (LFTF) comprising nanoencapsulated PCM capsules with *n*-Tetradecane as core and polystyrene (PS) made by ultrasonic-assisted miniemulsion polymerization with the *n*-Tetradecane core content of 89 wt.%. PS/*n*-Tetradecane nanocapsules with an average diameter of 132 nm were well distributed in the prepared slurry and formed the main energy storage components that possessed the melting and freezing temperatures of 4.04°C and −3.43°C along with their corresponding latent heat enthalpies values of 98.71 J/g and 91.27 J/g, respectively.

Fang et al. [93] synthesized polystyrene (PS)/*n*-dotriacontane (Dot) nanoencapsulated phase change material (NEPCM) by miniemulsion polymerization technique initiated ultrasonically and studied the influence of parameters such as time and temperature, ultrasonic power output, the ratio of co-emulsifiers used (sodium dodecyl sulfate (SDS) and poly-(ethylene glycol) monooctylphenyl ether (OP-10)), Styrene/Dot ratio, etc., on average diameter, rate of polymer conversion, encapsulation efficiency, and latent heat capacities of prepared NEPCM. The results suggested that the applied ultrasonic method produced nanocapsules of a mean diameter of 168.2 nm with high conversion rates (95.02%) and good encapsulation efficiency (61.23%) and high latent heat enthalpy of 174.8 J/g. Fu et al. [94] produced a nanoencapsulated PCM slurry for cold energy storage consisting of *n*-Tetradecane (Tet) core-based nanocapsules with hybrid shell wall (polystyrene-silica, PS-SiO_2), dispersed in a base fluid. Tet/PS-SiO_2 nanocapsules with an average size of 151.3 nm had a melting enthalpy of 83.38 KJ/kg. The incorporation of SiO_2 to nanocapsules by grafting SiO_2 on PS shells led to an increment in the thermal conductivity of Tet/PS-SiO_2 slurry by 8.4%.

Titanium dioxide (TiO_2) nanoparticles decorated (poly-(4-methylstyrene-co-divinylbenzene)) PCM microcapsules containing hexadecane were prepared following a one-pot non-Pickering emulsion templated suspension polymerization. Energy-dispersive X-ray (EDX) and X-ray diffraction (XRD) analysis confirmed the presence of TiO_2 over the polymeric shell. Metal oxide decorated PCM microcapsules with 2.6 wt.% TiO_2 potentially exhibited maximum latent heat enthalpy of 174 J/g with an encapsulation efficiency of 76.6%. Additionally, these microcapsules also showed excellent photo-catalytic activity by synergistic photo-thermal effect and can be utilized as solar energy storage and depollution materials [95]. A flexible composite material based on styrene butadiene styrene (SBS) as a supporting material, paraffin as a PCM, and EG as a thermal conductivity enhancer was fabricated by dissolving in an organic solvent and utilized in battery thermal management system. It was observed that during the 5°C discharge process, the maximum temperature of the battery module can be maintained below 46°C and the temperature difference can be controlled within 4°C. Also, after heating at 55°C for 5 hours, the quality of the composite remained above 99%. As a supporting elastic network material, SBS can effectively improve the flexibility and elastic properties of the PCM. The presence of EG significantly improves the thermal conductivity of the composite PCM, with the thermal conductivity of SBS/paraffin/EG (4%) reaching 0.88 W/(Km); however, the same PCM does not continue to increase when the weight fraction of EG exceeds 4% [96].

6.4. Acrylic polymer-based PCMs

Wang et al. [97] developed microencapsulated PCMs with the value-added function of thermo-chromic property by the introduction of thermo-chromic pigments to the acrylic polymer shells. The thermo-chromic pigments were first treated with a silane coupling agent to improve the affinity toward organic phases and then incorporated on the interface of *n*-octadecane droplets present in oil/water emulsion leading to thermochromic pigment/PMMA shell formation via suspension polymerization technique, as shown in **Figure 7**. Carreira et al. [98] studied the role of thermal initiators such as Benzoyl Peroxide (BPO), Azobisisobutyronitrile (AIBN), and Trigonox 23 (TRIG) and different reaction temperatures on the rate of polymerization and PCM microcapsule properties for the synthesis of octadecane encapsulated acrylic microcapsules, using suspension technique, for thermo-responsive textiles. The obtained microcapsules possessed both energy storage enthalpy, about 175 J/g, and good thermal stability. Huang et al. [99] synthesized and prepared a series of microencapsulated PCMs for thermo-regulated fabrics by selecting poly (allyl methacrylate) (PALMA) as the shell material for encapsulation of organic binary PCM cores including butyl stearate (BS) and tetradecanol (TD). PALMA microencapsulated binary PCMs with an encapsulation content of 76.4 wt.% exhibited phase-transition temperatures and latent energy values of 20.1–32.6°C and 73.3 J/g, respectively. Zhao et al. [100] used a suspension polymerization technique to synthesize microencapsulated PCMs with dual functions of UV-shielding and TES. The developed microcapsules comprised

octadecane as the core PCM confined in polymethylmethacrylate (PMMA) shell embedded with modified TiO_2 nanoparticles, as the UV absorber. DSC analysis showed that the melting and crystallization enthalpy and encapsulation efficiency of doped microcapsules were 48.26 J/g, 153.82 J/g, and 73.2%, respectively. The thermal cycling stability of the synthesized PCM is shown in **Figure 8**.

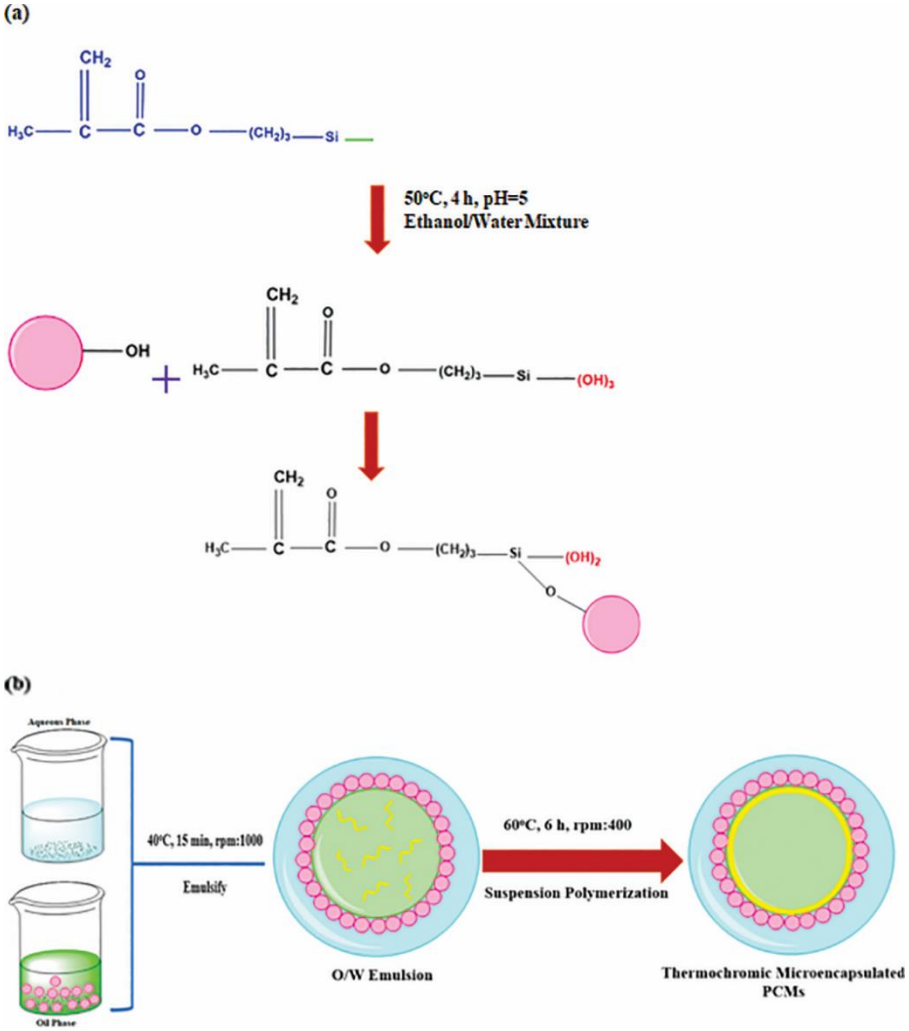


Figure 7 • Schematic fabrication process of thermo-chromatic microencapsulated phase change materials. (a) The modification process of thermo-chromic pigments. (b) The formation process of the TC-micro PCMs based on *n*-octadecane core and thermo-chromic pigments/PMMA shell through suspension-like polymerization. Adapted with permission from [97].

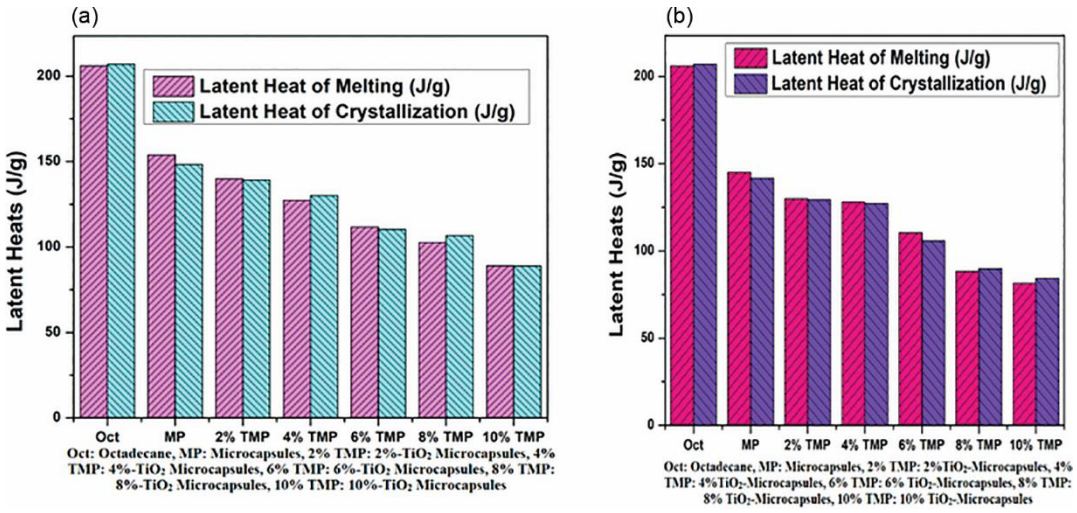


Figure 8 • Thermal cycling stability of synthesized microencapsulated PCMs with varying dosage of titanium nanoparticles (a) 1st cycle and (b) 500th cycle. Data from [100].

Wang et al. [101] obtained nanoencapsulated PCMs by encapsulating stearic acid inside PMMA nanocapsules by using an ultrasonically initiated in-situ polymerization method. DSC data suggested that the thermal storage ability and encapsulation

efficiency of fabricated PMMA/stearic acid nanocapsules were 155.6 J/g and 83.0%, respectively. Su et al. [102] produced microencapsulated PCMs following the free-radical polymerization route where they opted for *n*-octadecane as a core latent heat storage material and PMMA as a shell component, which exhibited a phase change at 22.68°C and energy storage density of 170 kJ/kg with thermal stability behavior above 160°C. Şahan et al. [103] used acrylic copolymer shells to successfully encapsulate behenic acid (BeA) via emulsion polymerization technique to obtain microcapsules with uniform spherical geometry, with particle size ranging below 500 nm. They showed a phase change temperature ranging between 65°C and 85°C with maximum latent heat capacity of 58 J/g. The resulted product was reported suitable for potential applications such as industrial waste heat, electronics, solar residential heating, lithium-ion batteries, automotive applications, and so on.

6.5. Polyamide polymer-based PCMs

Cai et al. [104] determined the structural morphologies and TES properties of electrospun polyamide 6 nanofibers-based phase change composites containing fatty acid eutectics, mainly capric acid (CA). The fabricated composite nanofibers had phase change temperatures and corresponding enthalpies lower than that of pure CA. Cai et al. [105] studied the impact of 10 wt.% of EG on structural and thermal properties of Nylon-6 electrospun nanofibrous mats consisting of fatty acid eutectic mixtures of CA. These polyamide nanofibers composites were made as PCMs by soaking them for ten hours until saturated absorption into an ultrasonically melt-blended solutions of capric acid (CA)-lauric acid (LA), capric acid (CA)-palmitic acid (PA), and capric acid (CA)-stearic acid (SA) to allow their absorption by the fabricated nanofibrous mats. SEM images indicated that Nylon-6/EG nanofibers become more uniform and straight upon EG addition; also, the average diameter of Nylon-6/EG was lower than pure Nylon-6. Casado et al. [106] worked on the design optimization and improvement of fatigue resistance of injection-molded short-fiber glass-reinforced polyamide structural parts used in high-speed rail fastenings. The addition of an inorganic hydrated salt PCM ($Na_2S_2O_3.5H_2O$) provided a cooling effect to the reinforced polyamide component and controlled the increase in polymer temperature, as a viscoelastic response under the dynamic cyclic fatigue loads. Fatigue resistance of these structural components filled with PCMs increased by 400% against the polyamide parts without PCMs.

6.6. Polyester polymer-based PCMs

Cai et al. [107] investigated the outcome of the silica nanoparticles (nano-SiO_2) addition on morphological, thermal stability, heat storage, and combustion characteristics of electrospun LA containing polyethylene terephthalate (PET) composite fibers. DSC analysis indicated that the heat storage enthalpy values initially increase with increasing nano-SiO_2 and then decrease with further nanoparticles loading. Golestaneh et al. [108] fabricated phase change composite fibers by an innovative co-electrospinning method utilizing eutectics mixtures of binary fatty acids: capric-lauric acid (CA-LA) and capric-palmitic acid (CA-PA) as the energy storage component and PET as the supporting matrix fibers (**Figure 9**). Morphological results indicated that the fatty acids were fairly dispersed in the PET filaments and the mean diameter of composite fibers improved by raising the PCM concentration in the PET matrix. The fabricated fibers possessed a wider range of phase change temperatures (−9 to 34°C).

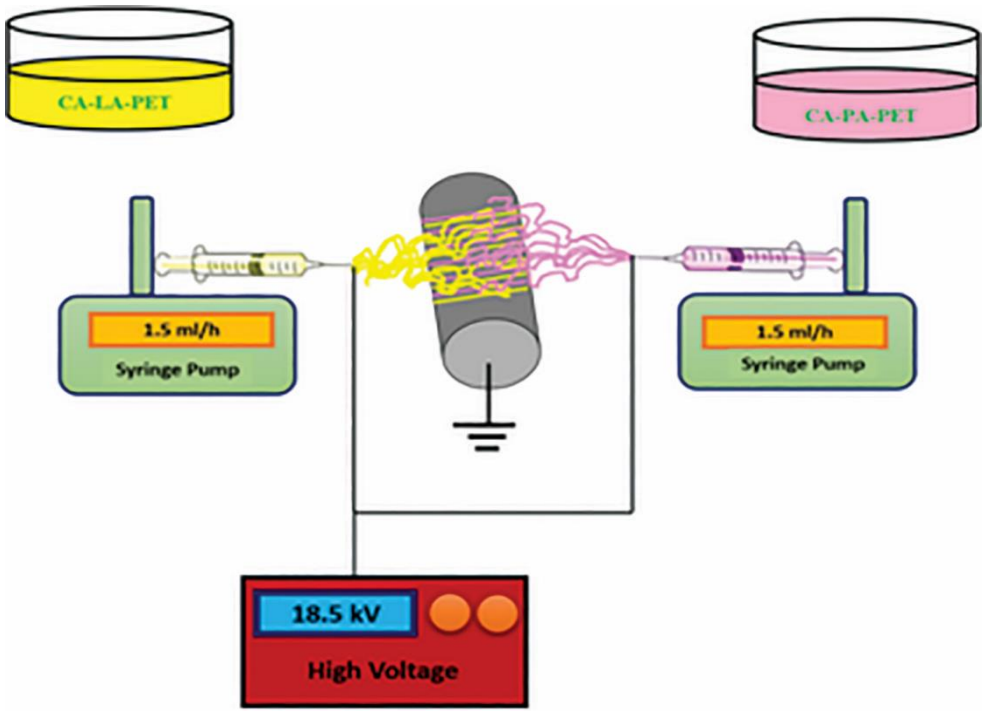


Figure 9 • The schematic view of the co-electrospinning setup. Adapted with permission from [108].

Nejman et al. [109] studied the effects of heating/cooling cycle rates (1, 5, 10, 15°C/min) on TES and retrieval properties of polyester-based knitted and woven fabrics impregnated with polymer pastes, containing 40 wt.% of two microencapsulated PCMs. It was evident that an increase in heating/cooling rates led to an increase in endothermic phase transitions and a decrease in crystallization temperatures. The variations of latent heat energies of fabrics with different heating rates are illustrated in **Figure 10**.

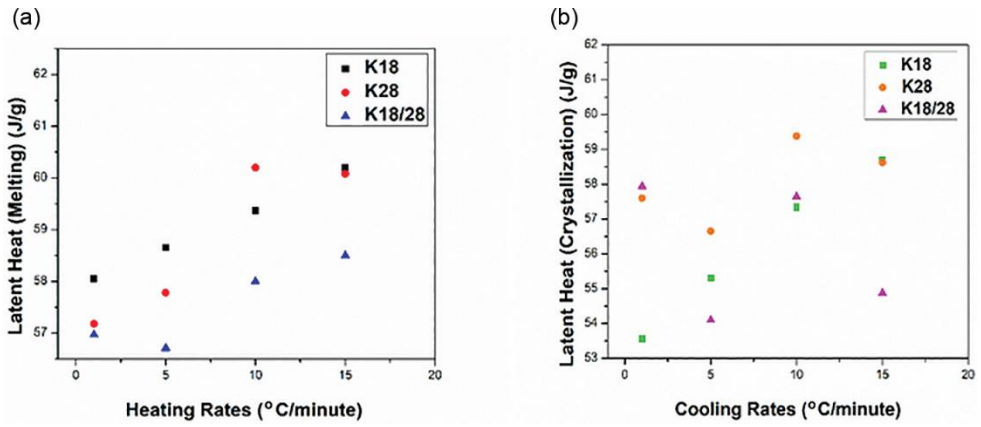


Figure 10 • Phase change temperatures and variation of latent heat energies of fabrics with different heating rates. (a) Latent heat of melting and (b) latent heat of crystallization. Data from [109].

6.7. Aminoplastic-based PCMs

Yoo et al. [110] investigated the effect of incorporation of eicosane-filled melamine-formaldehyde microcapsules into glass fibers (GF) reinforced epoxy resin on thermo-physical properties of fiber-reinforced plastics composites. Thermal conductivities of produced composites showed decreasing trends with increasing microencapsulated PCMs, owing to the low thermal conductivity of capsule shell layer. Thermal stabilities of the prepared composite were decreased due to lack of interfacial adhesion between fibrous filler and the matrix. Hu et al. [111] used an in-situ polymerization technique to achieve carboxymethylcellulose (CMC)-modified melamine-formaldehyde (MF) spherical nanocapsules of an average diameter of around 50 nm. The DSC test reported that the heat energy enthalpies were improved by CMC addition to a certain concentration and reached an optimum latent heat enthalpy value of 83.46 J/g, with an encapsulation ratio of confined paraffin of 63.1%. Zhang et al. [112] incorporated graphene oxide (GO) nanosheets as an extra protective layer at the junction between core and shell materials made by in-situ polymerization composed of paraffin and melamine-formaldehyde resin (MF) to develop a new type of microencapsulated PCMs as represented in **Figure 11**. The average diameters of the capsules gradually enlarged with an upsurge in GO content from 6.32 to 15.89 µm against a diameter of 5.63 µm in the absence of GO nanosheets. An additional GO layer enabled the prepared microcapsules to efficiently control and minimized the paraffin leakage rate from the MF shell by 93.1% during 50 hours of thermal exposure.

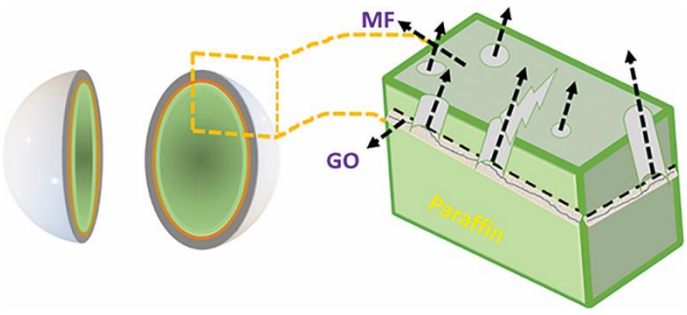


Figure 11 • Schematic view for possible permeation through the shells with GO. Adapted with permission from [112].

Konuklu et al. [113] used various aminoplastic-based shell materials including urea-formaldehyde (UF) resin, melamine-formaldehyde (MF) resin, and urea-melamine-formaldehyde resin for encapsulation of CA, by coacervation method for TES purposes. It was clear that UF resin was a more suitable capsule material among all other used shell materials which offered latent heat enthalpies of 93.9 and 106.1 J/g during the heating and cooling process, respectively. Liu et al. [114] developed a LFTF to be employed as direct absorption solar collectors (DASCs) by dispersing the paraffin encapsulated graphite nanoparticles modified MF microcapsules in an ionic liquid, 1-butyl-3-methylimidazolium tetrafluoroborate ([BMIM]BF4). The microencapsulated slurries with encapsulated paraffin content of 51.1 wt.% exhibited phase change temperature and latent heat enthalpy of 50.5°C and 90.8 J/g, respectively, while the incorporation of graphite nanoparticles to the MF shell enhanced the thermal conductivity of produced LFTF as illustrated in **Figure 12**. Upon irradiation, the temperature of LFTF increased from 30°C to 113°C proving the high photo-thermal performance and thermal storage potential of these developed slurries, which was about twice that of the pure ionic liquid.

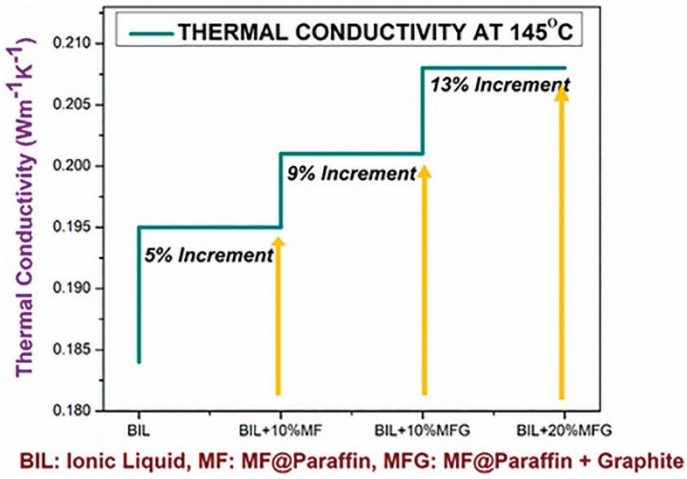


Figure 12 • Thermal conductivity of [BMIM]BF4 and the slurries containing different microencapsulated PCMs. Data from [114].

Microencapsulated PCMs composed of graphene oxide supported urea-formaldehyde (UF) resin capsules, filled with paraffin wax with a phase change temperature of 30.83°C and an associated heat of fusion equal to 90.23 J/g were developed via in-situ polymerization technique. This solution was further mixed with cement to form a cement slurry to control the heat evolution and prevent or delay the undesirable speed of the

hydration reaction in the cement, in addition to reducing the contents of C–S–H and $Ca(OH)_2$. The thermal conductivity values were found to be varied linearly with increasing order of graphene oxide concentrations where at 0.30% graphene oxide content, a hike of 53.84% in the thermal conductivity value was observed. This evaluated cement slurry can be resourceful in meeting the requirements of marine sector [115]. Gao et al. [116] focused on the synthesis of PCM slurry for direct absorption solar collector with optimal heat storage properties and photo-thermal conversion, by dispersing the spherical PCM microcapsules of 4 μm in diameter made up of melamine-formaldehyde (MF) resin and octadecane containing oleic acid-coated magnetic nanoparticles (OA-MNs) as the shell and core, respectively, into MWCNTs-H_2O nano-fluids. The prepared super-paramagnetic micro-encapsulated PCM particles had encapsulation efficiency of 63.51% corresponding to phase change enthalpy of 145 J/g. The oleic acid-modified magnetic nanoparticles were found to have boosted the thermal conductivity of the PCM microcapsules. Therefore, the resultant slurry consisting of 0.01 wt.% MWCNTs and 15 wt.% PCMs had the optimal photo-thermal conversion properties and thermal storage capacity.

Srinivasaraonaik et al. [117] used an in-situ polymerization technique to obtain the microencapsulated PCM consisting of fatty acid eutectic mixture—75% stearic acid (SA) + 25% CA as the core and MF resin as the shell material. They found out that a balanced pH (3.2) and stirring rates (1,500 rpm) controlled the microstructures of the obtained microencapsulated PCM, having spherical configurations with the smooth surface morphology and a particle size of 10.41 μm. The melting point and latent heat of fusion of the produced microcapsules were found to be 34.5°C and 103.9 kJ/kg, respectively. The obtained microencapsulated PCM also showed improved thermal stability, because of the outer MF resin protective layer as compared to the raw PCM with high encapsulation efficiency of 85.3%. Zhang et al. [118] prepared zirconium carbide (ZrC) nanoparticles decorated with the microencapsulated PCM containing melamine-urea-formaldehyde (MUF) shells filled with paraffin and studied the effect of ZrC nanoparticles on microstructure, thermal storage, thermal stability, thermal conductivity, and photo-thermal conversion behavior. They observed that the thermal properties of microcapsules consisting of 4% ZrC nanoparticles were better than the unfilled microencapsulated PCM and concluded that the enthalpy was slightly increased by 14.92% and the thermal conductivity was increased by 225.16%, owing to the presence of metal nanoparticles. In addition, they showed remarkable optical absorption capability of 70.23%. Nanoencapsulated PCMs comprising of paraffin wax as latent heat storage core component and melamine-formaldehyde (MF) as shell were synthesized by the in-situ polymerization technique. The spherical nanocapsules were present in average diameter of 260–450 nm. They undergo phase transitions at 49°C and 54°C, respectively, with corresponding latent of fusion and crystallization recorded as 107.4–135.3 and 109.8–133.1 J/g, respectively. They possessed good thermal stability and also exhibited decent thermal reliability after withstanding 2,000 thermal cycles [119]. A series of microencapsulated PCMs with boron nitride (BN) reinforced melamine-formaldehyde resin shell for improving the thermal conductivity was successfully synthesized via in-situ polymerization technique. The obtained metal particles decorated with PCM microcapsules exhibited spherical profiles with a diameter ranging between 5 and 10 μm and possessed high latent heat storage density of 145.5 J/g. Moreover, highly thermally conductive BN effectively enhances the thermal conductivity of the microcapsules and supplies them with a fast thermal-responsive rate [120]. Naikwadi et al. [121] put efforts to develop solar TES coating using melamine-formaldehyde (MF) shell microcapsules with an *n*-Tetracosane as core PCM, which were subjected to an epoxy primer-polyurethane (EP-PU) topcoat system in different loading concentrations (0–25 wt.%). The microcapsules tend to melt at 53.07°C with latent heat of melting of 134.74 J/g and crystallize at 48.81 with latent heat crystallization of 133.32 J/g. The phase change properties of the coated material were found to increase with the increase in the microcapsule loading. Coating properties (pencil hardness, flexibility, and scratch resistance) deteriorate with increased microcapsule loading.

6.8. Polyurea and polyurethane polymers-based PCMs

Lu et al. [122] carried out the microencapsulation of BS inside a three-dimensional polyurethane (PU) shell network using pentaerythritol as a cross-linking agent, via interfacial polymerization. The resultant cross-linked microcapsules possessed smooth and compact surface morphology with good thermal stability and heat storage characteristics. Zhan et al. [123] synthesized polyurea microcapsules encompassing paraffin as a core for TES applications by interfacial polycondensation technique. The resultant polyurea microcapsules showed heat of phase change about 92.5 J/g with successful encapsulation of 44.5 wt.% of paraffin PCM. Sundararajan et al. [124] synthesized hyperbranched polyurethanes by reacting an adduct of PEG-IPDI capped isocyanate with a trifunctional monomer, trimethylolpropane (TMP) in varying molar concentrations, as a polymeric solid–solid PCM. The prepared polymeric solid–solid PCM possessed maximum fusion enthalpy of 145 J/g, which was found to increase with increasing PEG mass ratio and undergoes a phase transition at 55°C. Zhou et al. [125] incorporated HNTs as a reinforcing agent to the prepared solid–solid PCM made by reacting polyethylene glycol (PEG, M.W. 4000) and hexamethylene diisocyanate biuret (HDIB), leading to cross-linked polyurethane (PU) structures. HNTs offered excellent thermal stability. However, with an increase in HDIB concentrations, the phase change temperature decreased from 32.9°C to 24.2°C and the latent enthalpies reduced from 123.7 to 85.8 J/g owing to excessive cross-linking.

Wang et al. [126] utilized Pickering emulsion polymerization to synthesize polyurethane shells filled with methyl laurate-based PCMs. The obtained microcapsules exhibited core–shell arrangement and a smooth surface with shell thickness of 8~10 μm and the particle size ranging between 150 and 350 μm. The melting and crystallization enthalpies of microcapsules were as high as 147.71 and 143.36 J/g, respectively, accompanied with 83.34% of encapsulation efficiency. They also showed excellent thermal reliability and thermal cycling stability. Zhu et al. [127] developed temperature controlling coating materials containing microencapsulated PCMs composed of *n*-octadecane as core and polyurea as shell walls, which were synthesized via interfacial polymerization technique. Morphological studies reveal the PCM microcapsules can maintain its original shape and be dispersed uniformly in the coatings with 30 wt.% microencapsulated PCM loads and ensured the latent heat capacity of microencapsulated PCMs and the heat storage coatings are 151.89 and 45.5 J/g,

respectively. Test chambers having wall boards painted with and without microencapsulated PCM coatings were tested under the same conditions and compared. Overall, the chamber with microencapsulated PCM coatings can maintain a stable temperature of 26–28°C for approximately 77 minutes when the heating load is 80 W/m². However, the chamber without PCM coating can only maintain such temperature for 13 minutes.

7. Other non-polymeric-based PCMs

In a recent study, PCM composites were developed by using the eutectic PCMs mixture of CA and LA as the active PCM and waste fly ash as a supporting material. The performance of these PCM composites was evaluated through simulation, by taking Guangdong as a reference city, with desirable temperature around 25°C. The required temperature properties were achieved by using a mass ratio of CA/LA of 4:6. The optimum FA/PCMs composite has a latent heat of 45.38 J/g and has good thermal reliabilities. The results proved that the preparation of such PCMs is an efficient way for the reuse of solid waste FA by subjecting the inorganic material for good contribution to the development of energy-saving buildings [128]. Li et al. [129] fabricated PCM composites by introducing PEG into mesoporous pores of ZSM-5 assisted with vacuum impregnation. The latent heat of the as-prepared PCMs reaches 76.37 J/g without any leakage. Shape-stabilized PCMs based on polyethylene glycol (PEG) loaded in varying loadings of 60–90 wt.% confined in porous fumed silica (FS) as the porous support were successfully investigated. The optimum 80 wt.% PEG/FS PCM possessed a high crystallinity of 93.1%, corresponding to a remarkable TES capacity of 130.6 J/g, and excellent thermal reliability post experiencing 500 thermal cycles. It was observed that porous silica provided sufficient cavities to confine a large amount of PEG with interactions such as surface tension, capillary, and interfacial hydrogen bonds (H-bond). The interfacial H-bonds negatively affected the crystallinity of PEG and decreased the TES capacity, which could be relieved by a large content of confined PEG [130]. He et al. [131] successfully developed temperature-regulating cement-based PCMs by incorporating PCM microcapsules composed of CA as latent heat storage component and hydroxyapatite as the shell wall material. Fatty acid-filled hydroxyapatite microcapsules exhibited high CA absorption ratio and good physicochemical stability. The incorporation of CA/HAP in cement-based materials leads to a decrease in workability, acceleration of hydration process, and significant improvement of thermal regulation property. The compressive strength of mixtures containing PCM capsules is noted to be increased when incorporating 10% of PCM microcapsule by volume.

8. Conclusions

This comprehensive review presents the necessity of energy storage and known TES technology, including PCMs available to mankind. Fatty acid-based PCMs can be effectively useful in reducing the reliance on the latent heat storage materials generated through fossil fuels while potently acting as an environmentally friendly energy solution in making a “cleaner” and “greener” ecosystem as well as boosting the concept of sustainable development. This study will help readers in understanding the working, selection criteria, nature of a broad range of PCMs, and also storage methodologies for their effective usage in practical applications. This detailed literature study is focused on highlighting the significance of polymers in vast research and development activities carried out by various research groups in the pathway of search of future PCMs for latent heat storage technology based on polymers. The scattered research available on this significant topic has been clubbed up to shape up as a detailed review on the utility of polymers, whether natural or synthetic, in the development or production of new and advanced PCMs in the form of encapsulated or shape-stabilized PCMs composed of different types of PCMs, different shell wall materials, or supporting matrices.

Acknowledgments

The authors would like to thank Delhi Technological University for providing technical and financial assistance.

Funding

The authors declare no financial support for the research, authorship, or publication of this article.

Author contributions

Conceptualization, R.K.; methodology, S.T.; validation, R.K.; formal analysis, S.T. and R.K.; investigation, S.T.; resources, S.T.; data curation, S.T.; writing—original draft preparation, S.T.; writing—review and editing, R.K.; supervision, R.K.; project administration, R.K. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflict of interest.

Data availability statement

Data supporting these findings are available within the article, at https://doi.org/10.20935/AcadMatSci6184, or upon request.

Institutional review board statement

Not applicable.

Informed consent statement

Not applicable.

Sample availability

The authors declare no physical samples were used in the study.

Additional information

Received: 2024-01-22

Accepted: 2024-03-25

Published: 2024-05-02

Academia Materials Science papers should be cited as *Academia Materials Science 2024*, ISSN 2997–2027, https://doi.org/10.20935/AcadMatSci6184. The journal’s official abbreviation is *Acad. Mat. Sci.*

Publisher's note

Academia.edu stays neutral with regard to jurisdictional claims in published maps and institutional affiliations. All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright

© 2024 copyright by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

References

1. Dheep GR, Sreekumar A. Influence of nanomaterials on properties of latent heat solar thermal energy storage materials – a review. Energy Convers Manag. 2014;83:133.
2. Zhou D, Zhao CY, Tian Y. Review on thermal energy storage with phase change materials (PCMs) in building applications. Appl Energy. 2012;92:593.
3. Kant K, Shukla A, Sharma A. Advancement in phase change materials for thermal energy storage applications. Sol Energy Mater Sol Cells. 2017;172:82.
4. Liu C, Rao Z, Zhao J, Huo Y, Li Y. Review on nanoencapsulated phase change materials: preparation, characterization and heat transfer enhancement. Nano Energy. 2015;13:814.
5. Sharma RK, Ganesan P, Tyagi VV, Metselaar HSC, Sandaran SC. Developments in organic solid–liquid phase change materials and their applications in thermal energy storage. Energy Convers Manag. 2015;95:193.
6. Sharma A, Tyagi VV, Chen CR, Buddhi D. Developments in organic solid–liquid phase change materials and their applications in thermal energy storage. Renew Sustain Energy Review. 2009;13:318.
7. Schmit H, Frank M, Rathgeber C. Investigation of K2HPO4·6H2O as phase change slurry (PCS). Conference: 11th IIR Conference on Phase Change Materials and Slurries for Refrigeration and Air Conditioning, 2016. doi:10.18462/iir.pcm.2016.0026
8. Nagano K, Mochida T, Takeda S, Domanski R, Rebow M. Thermal characteristics of manganese (II) nitrate hexahydrate as a phase change material for cooling systems. Applied Thermal Eng. 2003;23:229–41.
9. Zhang X, et al. Calcium chloride hexahydrate/diatomite/paraffin as composite shape-stabilized phase-change material for thermal energy storage. Energy Fuels. 2018;32(1):916–21. doi: 10.1021/acs.energyfuels.7b02866
10. Luo C, Su Q, Mi W. Thermophysical properties and application of LiNO3-H2O working fluid. Int J Refrig. 2013;36:1689–700.
11. Rathod AC, Bandela CV, Rehman AR. Experimental study on phase change material based thermal energy storage system. Int Res J Eng Technol. 2017;4(11):1–6.
12. Akamo DO. Stabilization of low-cost phase change materials for thermal energy storage applications. IScience. 2023;26:107175; 1–17. doi: 10.1016/ j.isci.2023.107175
13. Schmit H, Pfeffer W, Rathgeber C, Hiebler S. Experimental investigation of the concentration dependent maximum storage capacity of two inorganic phase change materials. Energy Procedia. 2015;73:231–8.
14. Kumar N, Banerjee D, Chavez Jr R. Exploring additives for improving the reliability of zinc nitrate hexahydrate as a phase change material (PCM). J Energy Storage. 2018;20:153–62.
15. Shamberger PJ, Reid T. Thermophysical properties of potassium fluoride tetrahydrate from (243 to 348) K. J Chem Eng Data. 2013;58:294–300. doi: 10.1021/je300854w
16. Guna Singh KC, Hemalatha G. Impact of sodium silicate pentahydrate as phase change material in concrete cubes for enhancing the thermal comfort. Int J Civil Eng Technol. 2017;8(1):999–1007 A.
17. Chen W, et al. Macro-encapsulated 3D phase change material: $Na_2S_2O_3·5H_2O$-NaOAc·$3H_2O$/carbonized Melamine sponge composite as core and SiC modified polyurethane thin-layer as shell. Compos Sci Technol. 2021;214:108981.
18. Silalahi AO, et al. The enthalpy-temperature curve of new PCM $Ca(NO_3)_2·4H_2O$ and $Co(NO_3)_2·6H_2O$ for low temperature latent TES. J Phys Conf Ser. 2019;1204:012057.
19. Min KE, Jang JW, Kim JK, Wern C, Yi S. Thermophysical properties of inorganic phase change materials based on $MnCl_2·4H_2O$. Appl Sci. 2022;12(13):6338; 1–12.
20. Fu W, et al. Phase change temperature adjustment of $CH_3COONa·3H_2O$ to fabricate composite phase change material for radiant floor heating. Case Stud Therm Eng. 2023;42:102773.
21. Song Y, et al. Experimental and numerical investigation on dodecane/expanded graphite shape-stabilized phase change material for cold energy storage. Energy. 2019;189:116175.
22. Zhang H, et al. Preparation and thermal model of tetradecane/expanded graphite and a spiral wavy plate cold storage tank. Energies. 2022;15:9435. doi: 10.3390/en15249435
23. Catalina Velez C, Ortiz de Zarate JM, Khayet M. Thermal properties of n-pentadecane, n-heptadecane and n-nonadecane in the solid/liquid phase change region. Int J Therm Sci. 2015;94:139–46. doi: 10.1016/j.ijthermalsci.2015.03.001
24. Lashgari S, Arabi H, Mahdavian AR, Ambrogi V. Thermal and morphological studies on novel PCM microcapsules containing n-hexadecane as the core in a flexible shell. Appl Energy. 2017;190:612–22.
25. Can A, Žigon J. n-Heptadecane-impregnated wood as a potential material for energy-saving buildings. Forests. 2022;13:2137. doi: 10.3390/f13122137

26. Bui TH, Nguyen GT. Enhanced thermal energy storage of n-octadecane-impregnated mesoporous silica as a novel shape-stabilized phase change material. ACS Omega. 2022;7:12222–30.
27. Sarı A, Alkan C, Biçer A, Altuntas A, Bilgin C. Micro/nanoencapsulated n-nonadecane with poly(methyl methacrylate) shell for thermal energy storage. Energ Convers Manag. 2014;86:614–21.
28. Yang S, Huang B, Wang J, Lund PD. Characteristics of natural convection in n-eicosane in a square cavity with discrete heat source. Case Stud Therm Eng. 2021;27: 101245.
29. Sarı A, Alkan C, Biçer A. Thermal energy storage characteristics of micro-nanoencapsulated heneicosane and octacosane with poly(methylmethacrylate) shell. J Microencapsul. 2016;33:221–8, 1–8. doi: 10.3109/02652048.2016.1144820
30. Ho CJ, Jheng SR, Yang TR, Pourfattah F, Yan WY. Thermophysical properties of water-based nano-emulsion of tricosane - an experimental investigation. Case Stud Therm Eng. 2021;24:100849, 1–7.
31. Yamada Y, Horibe A. Thermal properties and related core/shell structure of n-tetracosane microencapsulated by calcium carbonate. Appl Therm Eng. 2020;178:115512. doi: 10.1016/j.applthermaleng.2020.115512
32. Vennapusa JR, Konala A, Dixit P, Chattopadhyay S. Caprylic acid based PCM composite with potential for thermal buffering and packaging applications. Mater Chem Phys. 2020;253:123453.
33. Pielichowska K, Pielichowski K. Phase change materials for thermal energy storage. Prog Mater Sci. 2014;65:67.
34. Fan Z, Zhao Y, Liu X, Shi Y, Jiang D. Thermal properties and reliabilities of lauric acid-based binary eutectic fatty acid as a phase change material for building energy conservation. ACS Omega. 2022;7:16097–108.
35. Zhou D, Yuan J, Zhou Y, Liu Y. Preparation and characterization of myristic acid/expanded graphite composite phase change materials for thermal energy storage. Sci Rep. 2020;10:10889.
36. Zhou D, Zhou Y, Yuan J, Liu Y. Palmitic acid-stearic acid/expanded graphite as form-stable composite phase-change material for latent heat thermal energy storage. J Nanomater. 2020;2020:1648080, 1–9.
37. Vennapusa JR, et al. Thermal buffering performance of a propyl palmitate/expanded perlite-based form-stable composite: experiment and numerical modeling in a building model. Energy Fuels. 2021;35:2704–16.
38. Alkhazaleh AH. Preparation and characterization of isopropyl palmitate/expanded perlite and isopropyl palmitate/nanoclay composites as form-stable thermal energy storage materials for buildings. J Energy Storage. 2020;32: 101679.
39. Chinnasamy V, Appukuttan S. Thermo-physical investigation of butyl stearate as potential phase change material for thermal energy storage in cooling application. Energy Storage. 2020;160:1–9. doi: 10.1002/est2.160
40. Chi B, Yan Yao Y, Cui S. Preparation and numerical simulation of heat transfer performance for methyl palmitate/expanded graphite phase change composite. Key Eng Mater. 2020;834:132–8. doi:10.4028/www.scientific.net/KEM.834.132
41. Tabassum H, Huang X, Chen R, Zou R. Tailoring thermal properties via synergistic effect in a multifunctional phase change composite based on methyl stearate. J Materiomics. 2015;1(3):229–35.
42. Liston LC, et al. Binary mixtures of fatty acid methyl esters as phase change materials for low temperature applications. Appl Therm Eng. 2016;96:501–7.
43. Philip N, Dheep GR, Sreekumar A. Cold thermal energy storage with lauryl alcohol and cetyl alcohol eutectic mixture: thermophysical studies and experimental investigation. J Energy Storage. 2020;27:101060, 1–15.
44. Wang C, Wang T, Hu Z, Cai Z. Facile synthesis and thermal performance of cetyl palmitate/nickel foam composite phase change materials for thermal energy storage. J Energy Storage. 2020;28:101179, 1–7.
45. Menon AM, Anirudh K, Abhilash A, Aby Kurian M. Comparison of a conventional domestic refrigerator with a PCM encapsulated refrigerator. Int J Eng Res Technol. 2015;4(10):1–6.
46. Goliya S, Gandhi S, Murthy YI. Strength characteristics of concrete containing glycerine as phase change material. J Adv Eng Res. 2023;10(1):42–7.
47. Budiman AC, et al. Phase change material composite battery module for thermal protection of electric vehicles: an experimental observation. Energies. 2023;16:3896.
48. Amberkar T, Mahanwar P. Study of beeswax–carbon nanotubes– paperboard nanocomposite for temperature regulation of packaging container. Mater Proc. 2022;9:9. doi: 10.3390/materproc2022009009
49. Brahma B, Narzary R, Baruah DC. Acetamide for latent heat storage: thermal stability and metal corrosivity with varying thermal cycles. Renew Energy. 2020;145:1932–40.
50. Purohit K, Murty VVS, Dixit RC, Sharma A. Development of an acetanilide/benzoic acid eutectic phase change material based thermal energy storage unit for a passive water heating system. Bull Mater Sci. 2019;42:119, 1–9.
51. Miro L, et al. Health hazard, cycling and thermal stability as key parameters when selecting a suitable phase change material (PCM). Thermochim Acta. 2016;627–629:39–47.
52. Cabeza LF, Castell A, Barreneche C, de Gracia A, Fernandez AI. Materials used as PCM in thermal energy storage in buildings: a review. Renew Sustain Energy Rev. 2001;12: 1675–95.
53. Mehling H, Cabeza LF. Heat and cold storage with PCM. An up to date introduction into basics and applications. Berlin/Heidelberg: Springers-Verlag; 2008.
54. Cabeza LF, Mehling H, Hiebler S, Ziegler F. Heat transfer enhancement in water when used as PCM in thermal energy storage. Appl Therm Eng. 2002;22:1141–51.

55. Abhat A. Low temperature latent heat thermal energy storage: heat storage materials. Sol Energy. 1983;30:313–32.

56. Khudhair AM, Farid MM. A review on energy conservation in building applications with thermal storage by latent heat using phase change materials. Energy Convers Manage. 2004;45:263–75.

57. Sari A, Karaipekli A. Preparation and thermal properties of capric acid/palmitic acid eutectic mixture as a phase change energy storage material. Mater Lett. 2008;62:903–6.

58. Keles S, Kaygusuz K, Sari A. Lauric and myristic acids eutectic mixture as phase change material for low-temperature heating applications. Int J Energy Res. 2005; 29:857–70.

59. Sari A. Eutectic mixtures of some fatty acids for low temperature solar heating applications: thermal properties and thermal reliability. Appl Therm Eng. 2005;25:2100–7.

60. Sari A, Kaygusuz K. Thermal performance of a eutectic mixture of lauric and stearic acids as PCM encapsulated in the annulus of two concentric pipes. Sol Energy. 2002;72: 493–504.

61. Sari A. Thermal characteristics of a eutectic mixture of myristic and palmitic acids as phase change material for heating applications. Appl Therm Eng. 2003;23:1005–17.

62. Sari A, Kaygusuz K. Thermal energy storage characteristics of myristic and stearic acids eutectic mixture for low temperature heating applications. Chin J Chem Eng. 2006;14:270–5.

63. Baran G, Sari A. Phase change and heat transfer characteristics of a eutectic mixture of palmitic and stearic acids as PCM in a latent heat storage system. Energy Convers Manage. 2003;44:3227–46.

64. Lin Z, Wei Z, Lin Shuang L, Jin Xiangyu J. Experimental study on improving stability of PCM and MEPCM slurry with different surfactants. Int J Low-Carbon Technol. 2018;13:272–6.

65. Omara AAM, et al. An experimental study on using polyethylene glycol (PEG) 600 as phase change material for thermal comfort and energy saving in buildings. International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2018:1–6.

66. Niles N, Vijayarajah V. A study of the effect of PEG-1000 phase change materials coating on selected fabric properties. J Ind Text. 2022;52:1–12. doi:10.1177/15280837 221123066

67. Soodmand AM, Nejatbakhsh S, Pourpasha H, Hassan Aghdasinia H, Heris SZ. Simulation of melting and solidification process of polyethylene glycol 1500 as a PCM in rectangular, triangular, and cylindrical enclosures. Alexandria Eng J. 2022;61:8431–56.

68. Ansu AK, Pandya M, Sharma RK, Tripathi D. Experimental investigation of hybrid PCM polyethylene glycol with $Al_2 O_3$ and CuO nanoparticles. Res Sq. 2022; in review. doi: 10.21203/rs.3.rs-2037341/v1

69. Sanchez CD, Cuadri AA, Navarro FJ, Partal P. Formulation and processing of novel non-aqueous polyethylene glycol-in-silicone oil (o/o) phase change emulsions. Solar Energy Mater Solar Cells. 2021;221:110898,1–10.

70. Zhang J, et al. Shape stability of polyethylene glycol/ acetylene black phase change composites for latent heat storage. Adv Mater Sci Eng. 2018;2018:3954163, 1–9. doi: 10.1155/2018/3954163

71. Ansu AK, Pandya M, Sharma RK, Tyagi VV, Tripathi D. Prediction of thermal properties and reliability testing of binary eutectic mixture of polyethylene glycol 2000 and 10000 as phase change materials. ChemistrySelect. 2020;5:9745–57.

72. Kou Y, et al. Thermal analysis and heat capacity study of polyethylene glycol (PEG) phase change materials for thermal energy storage applications. J Chem Thermodynamics. 2019;128:259–74. doi: 10.1016/j.jct.2018.08.031

73. Praveen B, Suresh S. Experimental study on heat transfer performance of neopentyl glycol/CuO composite solid-solid PCM in TES based heat sink. Eng Sci Technol. 2018;21:1086–94.

74. Bo T, Cheng TZ, Bin LR, Gong MC, Nan ZJ. Thermodynamic investigation of a solid–solid phase change material: 2-Amino-2-methyl-1,3-propanediol by calorimetric methods. Energy Convers Manag. 2010;51:1905–10. doi:10.1016/ j.enconman.2010.02.021

75. Zhang N, et al. A novel solid–solid phase change material: pentaglycerine/expanded graphite composite PCMs. Adv Eng Mater. 2018;20:1800237, 1–9. doi: 10.1002/adem. 201800237

76. Zhai M, Zhang S, Sui J, Tian F, Lan XZ. Solid–solid phase transition of tris(hydroxymethyl)aminomethane in nanopores of silica gel and porous glass for thermal energy storage. J Therm Anal Calorim. 2017;129:957–64. doi: 10.1007/s10973-017-6223-6

77. Venkitaraj KP, Suresh S. Experimental study on thermal and chemical stability of pentaerythritol blended with low melting alloy as possible PCM for latent heat storage. Exp Therm Fluid Sci. 2017;88:73–87. doi: 10.1016/j.exptherm flusci.2017.05.018

78. Peng H, Wenhua Guo W, Li M, Feng S. Melting behavior and heat transfer performance of gallium for spacecraft thermal energy storage application. Energy. 2021;228: 120575, 1-14.

79. Alam TE, Dhau JS, Goswami DY, Stefanakos E. Macro-encapsulation and characterization of phase change materials for latent heat thermal energy storage systems. Appl Energy. 2015;154:92–101. doi: 10.1016/j.apenergy. 2015.04.086

80. Sarier N, Onder E. Organic phase change materials and their textile applications: an overview. Thermochim Acta. 2012;540:7–60. doi: 10.1016/j.tca.2012.04.013

81. Karthikeyan M, Ramachandran T. Review of thermal energy storage of micro- and nanoencapsulated phase change materials. Mater Res Innov. 2014;18:541–54. doi: 10.1179/1433075X13Y.0000000143

82. Mu M, et al. Shape stabilised phase change materials based on a high melt viscosity HDPE and paraffin waxes. Appl

Energy. 2016;162:68–82. doi: 10.1016/j.apenergy.2015.10.030

83. Guo X, Cao J, Peng Y, Liu R. Incorporation of microencapsulated dodecanol into wood flour/high-density polyethylene composite as a phase change material for thermal energy storage. Mater Design. 2016;89:1325–34. doi: 10.1016/j.matdes.2015.10.068

84. Huang X, Alva G, Liu L, Fang G. Microstructure and thermal properties of cetyl alcohol/high density polyethylene composite phase change materials with carbon fiber as shape-stabilized thermal storage materials. Appl Energy. 2017;200:19–27. doi: 10.1016/j.apenergy.2017.05.074

85. Tang Y, Lin Y, Jia Y, Fang G. Improved thermal properties of stearyl alcohol/high density polyethylene/expanded graphite composite phase change materials for building thermal energy storage. Energy Build. 2017;153:41–9. doi: 10.1016/j.enbuild.2017.08.005

86. Tas CE, Unal H. Thermally buffering polyethylene/halloysite/phase change material nanocomposite packaging films for cold storage of foods. J Food Eng. 2021;292:110351, 1–8.

87. Premnath D, Vamsi VB, Tomar M. Freezing characteristics in finned LDPE spherical capsule. J Phys. 2021;2054:012053, 1-8. doi: 10.1088/1742-6596/2054/1/012053

88. Xie Y, et al. Paraffin/polyethylene/graphite composite phase change materials with enhanced thermal conductivity and leakage-proof. Adv Compos Hybrid Mater. 2021;4:543–51. doi: 10.1007/s42114-021-00249-6

89. Jin X, Li J, Xue P, Jia M. Solar Energy materials & solar cells preparation and characterization of PVC-based form-stable phase change materials. Sol Energy Mater Sol Cells. 2014;130:435–41. doi: 10.1016/j.solmat.2014.07.013

90. Zhang H, et al. Solar Energy Materials & Solar Cells A novel thermal-insulating film incorporating microencapsulated phase-change materials for temperature regulation and nano-TiO2 for UV-blocking. Sol Energy Mater Sol Cells. 2015;137:210–18. doi: 10.1016/j.solmat.2015.02.018

91. Giro-Paloma J, Konuklu Y, Fernández AI. Preparation and exhaustive characterization of paraffin or palmitic acid microcapsules as novel phase change material. Sol Energy. 2015;112:300–9. doi: 10.1016/j.solener.2014.12.008

92. Fang Y, et al. Preparation and thermal performance of polystyrene/n-tetradecane composite nanoencapsulated cold energy storage phase change materials. Energy Convers Manag. 2013;76:430–6. doi: 10.1016/j.enconman.2013.07.060

93. Fang Y, et al. Ultrasonic synthesis and characterization of polystyrene/n-dotriacontane composite nanoencapsulated phase change material for thermal energy storage. Appl Energy. 2014;132:551–6. doi: 10.1016/j.apenergy.2014.06.056

94. Fu W, et al. Thermo-physical properties of n-tetradecane@polystyrene-silica composite nanoencapsulated phase change material slurry for cold energy storage. Energy Build. 2017;136:26–32. doi: 10.1016/j.enbuild.2016.12.001

95. Parvate S, et al. Titanium dioxide nanoparticle-decorated polymer microcapsules enclosing phase change material for thermal energy storage and photocatalysis. ACS Appl Polym. Mater.,2021;3:1866–79. doi: 10.1021/acsapm.0c01410

96. Huang Q, Li X, Zhang G, Deng J, Wang C. Thermal management of Lithium-ion battery pack through the application of flexible form-stable composite phase change materials. Appl Thermal Eng. 2021;183:116151, 1–15. doi: 10.1016/j.applthermaleng.2020.116151

97. Wang H, et al. Fabrication and characterization of microcapsulated phase change materials with an additional function of thermochromic performance. Sol Energy. 2016;139:591–8. doi: 10.1016/j.solener.2016.10.011

98. Carreira AS, et al. Preparation of acrylic based microcapsules using different reaction conditions for thermoregulating textiles production. Eur Polym J. 2017;93:33–43. doi: 10.1016/j.eurpolymj.2017.05.027

99. Huang M, et al. Preparation and characterization of microencapsulated phase change materials with binary cores and poly (allyl methacrylate) (PALMA) shells used for thermo-regulated fibers. Thermochim Acta. 2017;655:262–8. doi: 10.1016/j.tca.2017.07.006

100. Zhao J, et al. Microencapsulated phase change materials with TiO2-doped PMMA shell for thermal energy storage and UV-shielding. Sol Energy Mater Sol Cells. 2017;168:62–8. doi: 10.1016/j.solmat.2017.04.014

101. Wang G, Xu W, Hou Q, Guo S.A simple sonochemical method for fabricating poly (methyl methacrylate)/stearic acid phase change energy storage nanocapsules. Ultrason Sonochem. 2015;27:403–7. doi: 10.1016/j.ultsonch.2015.06.007

102. Su W, Zhou T, Li Y, Lv Y. Development of microencapsulated phase change material with poly (methyl methacrylate) shell for thermal energy storage. Energy Procedia. 2019;158:4483–8.

103. Şahan N, Paksoy H. Designing behenic acid microcapsules as novel phase change material for thermal energy storage applications at medium temperature. Int J Energy Res. 2020;44(5):3922–33.

104. Cai Y, et al. Structural morphology and thermal performance of composite phase change materials consisting of capric acid series fatty acid eutectics and electrospun polyamide6 nanofibers for thermal energy storage. Mater Lett. 2012;89:43–6. doi: 10.1016/j.matlet.2012.08.067

105. Cai Y, et al. In fluences of expanded graphite on structural morphology and thermal performance of composite phase change materials consisting of fatty acid eutectics and electrospun PA6 nano fibrous mats. Renew Energy. 2013;57:163–70. doi: 10.1016/j.renene.2013.01.044

106. Casada JA, et al. Fatigue behavior enhancement of short fiber glass reinforced polyamide by adding phase change materials. Compos Part B Eng. 2016;93:115–22. doi: 10.1016/j.compositesb.2016.02.059

107. Cai Y, et al. Effects of nano-SiO2 on morphology, thermal energy storage, thermal stability, and combustion properties

of electrospunlauric acid/PET ultrafine composite fibers as form-stable phase change materials. Appl Energy. 2011; 88:2106–12. doi: 10.1016/j.apenergy.2010.12.071

108. Golestaneh SI, et al. Fabrication and characterization of phase change material composite fibers with wide phase-transition temperature range by co-electrospinning method. Appl Energy. 2016;182:409–17. doi: 10.1016/j.apenergy.2016.08.136

109. Nejman A. The impact of the heating/cooling rate on the thermoregulating properties of textile materials modified with PCM microcapsules. Appl Therm Energy. 2017;127: 212–23. doi: 10.1016/j.applthermaleng.2017.08.037

110. Yoo S, et al. Thermophysical properties of multifunctional glass fibre reinforced polymer composites incorporating phase change materials. Thermochim Acta. 2016;642:25–31. doi: 10.1016/j.tca.2016.09.003

111. Hu X, Huang Z, Zhang Y. Preparation of CMC-modified melamine resin spherical nano-phase change energy storage materials. Carbohydr Polym. 2014;101:83–8. doi: 10.1016/j.carbpol.2013.08.091

112. Zhang L, et al. Graphene oxide-modified microencapsulated phase change materials with high encapsulation capacity and enhanced leakage-prevention performance. Appl Energy. 2017;197:354–63. doi: 10.1016/j.apenergy.2017.04.041

113. Konuklu Y, Unal M, Paksoy HO. Microencapsulation of caprylic acid with different wall materials as phase change material for thermal energy storage. Sol Energy Mater Sol Cells. 2014120:536–42. doi: 10.1016/j.solmat.2013.09.035

114. Liu J, Chen L, Fang X, Zhang Z. Preparation of graphite nanoparticles-modified phase change microcapsules and their dispersed slurry for direct absorption solar collectors. Sol Energy Mater Sol Cells. 2017;159:159–66. doi: 10.1016/j.solmat.2016.09.020

115. Huo JH, Peng ZG, Yang SS. Novel cement slurry containing microencapsulated phase change materials: characterization, application and mechanism analysis. Int J Energy Res. 2020;44(6):4235–48.

116. Gao G, Zhang T, Guo C, Jiao S, Rao S. Photo-thermal conversion and heat storage characteristics of multi-walled carbon nanotubes dispersed magnetic phase change microcapsules slurry. Int J Energy Res. 2020;44(8):6873–84.

117. Srinivasaraonaik B, Singh LP, Tyagi I, Rawat A, Sinha S. Microencapsulation of a eutectic PCM using in situ polymerization technique for thermal energy storage. Int J Energy Res. 2020;44(5):3854–64.

118. Zhang Z, Fang J, Mu J. ZrC nanoparticle-modified microencapsulated phase change materials with enhanced thermal conductivity and photo-thermal conversion performance. Int J Energy Res. 2020;44(9):7283–98.

119. Zhang N, Yuan Y. Synthesis and thermal properties of nanoencapsulation of paraffin as phase change material for latent heat thermal energy storage. Energy Built Environ. 2020;1:410–6. doi: 10.1016/j.enbenv.2020.04.003

120. Xia Y, et al. Design and synthesis of novel microencapsulated phase change materials with enhancement of thermal conductivity and thermal stability: self-assembled boron nitride into shell materials. Colloids Surf. 2020;586:1–42. doi: 10.1016/j.colsurfa.2019.124225

121. Naikwadi AT, Samui AB, Mahanwar PA. Melamine-formaldehyde microencapsulated n-Tetracosane phase change material for solar thermal energy storage in coating. Sol Energy Mater Sol Cells. 2020;215:110676. doi: 10.1016/j.solmat.2020.110676

122. Lu S, et al. Preparation and characterization of cross-linked polyurethane shell microencapsulated phase change materials by interfacial polymerization. Mater Lett. 2018;211: 36–9. doi: 10.1016/j.matlet.2017.09.074

123. Zhan S, Chen S, Chen L, Hou W. Preparation and characterization of polyurea microencapsulated phase change material by interfacial polycondensation method. Powder Technol. 2016;292:217–22. doi: 10.1016/j.powtec.2016.02.007

124. Sundararajan S, Samui AB, Kulkarni PS. Synthesis and characterization of poly (ethylene glycol) (PEG) based hyperbranched polyurethanes as thermal energy storage materials. Thermochim Acta. 2017;650:114–22. doi: 10.1016/j.tca.2017.02.011

125. Zhou Y, et al. Synthesis and properties of crosslinking halloysite nanotubes/polyurethane- based solid-solid phase change materials. Sol Energy Mater Sol Cells. 2018;174: 84–93. doi: 10.1016/j.solmat.2017.08.031

126. Wang Z, Ma W, Hua D, Wu L. Synthesis and characterization of microencapsulated methyl laurate with polyurethane shell materials via interfacial polymerization in pickering emulsions. Colloids Surf. 2020;600:124958. doi: 10.1016/j.colsurfa.2020.124958

127. Zhu X, Sheng X, Li J, Chen, Y. Thermal comfort and energy saving of novel heat-storage coatings with microencapsulated PCM and their application. Energy Build. 2021;251: 111349. doi: 10.1016/j.enbuild.2021.111349

128. Ji R, Zou Z, Liu L, Wei S, Qu S. Development and energy evaluation of phase change material composite for building energy-saving. Int J Energy Res. 2019;43(14):8674–83.

129. Li C, Yu H, Song Y, Zhao M. Synthesis and characterization of PEG/ZSM-5 composite phase change materials for latent heat storage. Renew Energy. 2018;121:45–52. doi: 10.1016/j.renene.2017.12.089

130. Nguyen GT. Polyethylene glycol/fumed silica composites as shape-stabilized phase change materials with effective thermal energy storage. RSC Adv. 2023;13:7621–31. doi: 10.1039/d2ra08134b

131. Wang Y, et al. Preparation and characterizations of hydroxyapatite microcapsule phase change materials for potential building materials. Cons Build Mater. 2021;297:1–8. doi: 10.1016/j.conbuildmat.2021.123576

Signal, Image and Video Processing
https://doi.org/10.1007/s11760-024-03376-x

ORIGINAL PAPER

Shallowfake and deepfake image manipulation localization using noise and RGB-based dual branch method

Deepak Dagar[1] · **Dinesh Kumar Vishwakarma**[1]

Received: 21 April 2024 / Revised: 8 June 2024 / Accepted: 12 June 2024
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2024

Abstract
The reliability of multimedia is being progressively tested by sophisticated Image Manipulation localization (IML) methods, which has led to the creation of the IML domain. A good manipulation model requires extracting non-semantic differences features between manipulated and authentic regions to exploit artifacts, which calls for explicit comparisons between the two areas. Existing models either use handcrafted-based feature methods, convolutional neural networks (CNNs), or a combination of both. Handcrafted feature methods assume the tampering beforehand, limiting their capabilities for diverse tampering operations, while CNNs model semantic information, which is not enough for the manipulation artifact. To improve these limitations, we have designed a dual-branch model that combines handcrafted feature noise and CNNs as an Encoder-decoder(ED) powered by the attention mechanism. This dual-branch model uses noise features on one branch and RGB on the other before feeding to an ED architecture for semantic learning and skip connection deployed to retain spatial information. Furthermore, this architecture uses channel spatial attention to strengthen further and refine the features' representation. Extensive experimentation on the shallowfakes dataset (CASIA, COVERAGE, COLUMBIA, NIST16) and deepfake datasets Faceforensics + + (FF + +) to demonstrate the superior feature extraction capabilities and performance to various baseline models with AUC score even reaching 99%. Also, it is one of the first methods to perform localization on the deepfake dataset. The model is relatively lighter, has 38 million parameters, and easily outperforms other State-of-the-Art(SoTA) models.

Keywords Image forensics · Manipulation localization · Noise inconsistencies · Deepfake localization · Shallowfakes and deepfake localization · Encoder-decoder

1 Introduction

In recent years, computer graphics and deep learning development have given people access to more sophisticated tools for creating false images. With the advancement of such sophisticated editing AI tools, it is possible to alter the content of multimedia data with great ease to produce highly realistic content. Therefore, images altered to transmit false information and propagate rumors will harm our lives, mainly if they do so quickly and widely throughout the Internet [1]. Traditional image-editing tools or graphics-based methods are called "shallow fakes" [2]. Deep learning-based methods, also called "Deepfakes", have replaced more conventional graphics-based methodologies in this methodology. Deepfake has gained attention as it allows the creation of highly realistic and believable content easily. As such sophisticated technologies advance, we require more advanced IML techniques to deal with manipulated images that are already out there and prevent security threats.

The localization task aims to identify altered regions in an image and to segment the manipulated regions at the pixel level. In shallow fakes, there are three types of image manipulation [2] (Fig. 1):

1) *Splicing* It is copying and pasting a section of one image into another.
2) *Copy-move* duplicating a section of an image.
3) *In-painting* removing portions of images and filling in the gaps with visually appealing information.

✉ Dinesh Kumar Vishwakarma
dvishwakarma@gmail.com

Deepak Dagar
deepakdagargate@gmail.com

[1] Biometric Research Lab, Department of Information Technology(IT), Delhi Technological University(DTU), Bawana Road, Delhi 110042, India

Published online: 25 June 2024

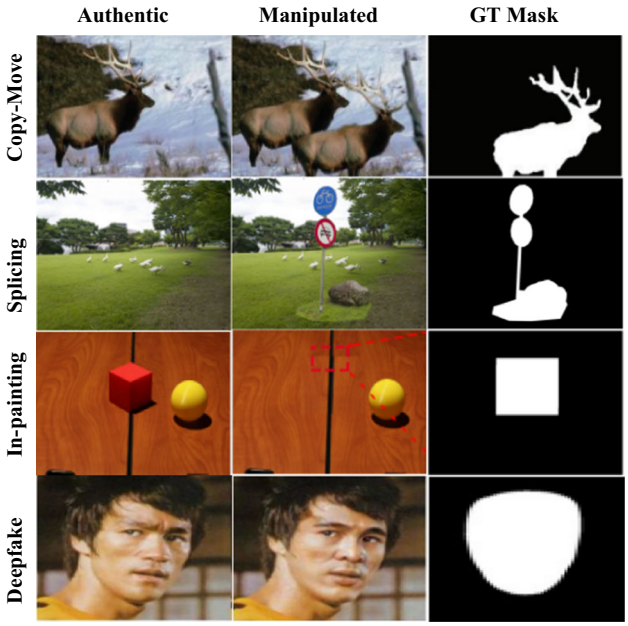


Fig. 1 Example of different kinds of image manipulation

Also, in Deepfake, three types of categorization can be made [1]:

1) *Face Swap* A face is transferred from the source image/frame to the target.
2) *Face Reenactment* Expression or movement is derived from the source image/frame to the target.
3) *Entire Image Synthesis* The powerful AI tools synthesize the entire image.

Handcrafted feature-based methods typically work assuming that a particular manipulation artifact exists. They seek to reveal the forgeries by investigating local discrepancies in color filter array [3], photo response non-uniformity noise (PRNU) [4], JPEG blocking artifacts [5], texture modules [6], and illumination [7]. However, these methods assume the tampering is known beforehand, limiting their capability in diverse scenarios. Deep learning methods, especially CNNs, generally tend to learn contextual(semantic) information and various SoTA methods of image forgery forensic techniques [8–10] resort to them [11]. However, CNNs have a locality basis, expound the relationship between neighboring regions, and may overlook the global manipulation between the artifacts. Inconsistencies in low-level features (like noise or high-frequency) that are not visible to the naked eye and distinctly visible traces are referred to as "artifacts" [12]. Accordingly, based on earlier research, the key to the IML task is to discover low-level inconsistencies and observable traces that are not semantic to capture the artifacts.

Handcrafted features have the limitation that they assume the tampering types beforehand and may overlook other kinds of tampering, and CNNs learn the semantic features and expound neighboring relationships for artifacts, which is not enough to learn the manipulation artifacts. To improve these limitations, we have designed a dual-branch model that uses handcrafted features such as noise inconsistencies on one branch and RGB information on another. This is one of the first few architectures to perform localization of the deepfake dataset. The following are the paper's primary contributions:

- Designed a novel dual-branch architecture that uses RGB information on one branch and noise features on another. Both branches employed channel and spatial attention, followed by ED architecture to represent artifacts better. A dual attention module for down-sampled features is used to learn semantic interdependencies along spatial and channel domains.
- Experimentation is done on the shallowfakes dataset, which is CASIA, Columbia, Coverage and NIST16, and the deepfake dataset FF + + outperforms other models to demonstrate the discriminative abilities of the model.
- Ablation studies are conducted to study the relevance of different components in the overall model.
- Complexity computation analysis was done to show that the model with 38 million parameters is relatively less complex than the various SoTA models.

2 Related work

Based on the feature representation, IML methods can be broadly classified into two groups: Handcrafted features and deep learning-based methods, which are discussed in this section.

2.1 Handcrafted features

Traditionally, by modelling authentic images, handcrafted feature-based techniques expose the statistical connection among pixels and capture the statistical deviation resulting from image tampering operations. For instance, the spatial rich model (SRM) [13], which is frequently used in image steganalysis, was generalized in [14] to extract residual-based features for the multidimensional Gaussian model and support vector machine (SVM) classifier for image manipulation detection and localization. However, altering an image will always result in changes to its visual components, which local image describers can use to detect forgeries. To detect the distortions by image splicing, [7] aggregated the statistical information retrieved by different local descriptors

that examine texture, lighting, shape, and color variables. Image manipulation affects the non-semantic features and results in low-level features like noise inconsistencies, and researchers expounded on such features to get better discriminative features. Zhang et al. [15] use an improved model of constrained convolution to extract noise features that serve as more subtle traces of manipulation fed to a dual-branch architecture for further feature learning. However, such noise features are inconsistent and robust regarding JPEG compression and Gaussian blur. Zhang et al. [16] also use an improved noise extraction method that uses two high-resolution and contextual branches to extract subtle manipulation traces. Handcrafted features assume that latent manipulation features are observable in specific manipulation artifacts, and a method designed to capture such artifacts would provide better results. However, this limits their generalization, and such methods are not robust enough for various manipulation operations.

2.2 Deep learning-based methods

Deep learning-based systems can automatically train and improve feature representations for forgery forensics, in contrast to conventional approaches that require the laborious feature engineering process to create the handcrafted features. Wu et al. [10] proposed that ManTra-Net detects and localizes generalized image manipulations. It recognizes the local anomalous features that can identify fabricated pixels and universal solutions to various forgeries. They are obviating the requirement for pre-and post-processing. However, the model has a few limitations: it cannot work for multi-forgeries and images with highly correlated noises. In Noiseprint [17], siamese networks were used to capture artifacts in the camera model based on noise residual features for manipulation localization. One of the main challenges in image manipulation detection is identifying specific generalizable features to avoid false alarms on authentic images yet sensitive to manipulations in novel data. Chen et al. [18] address both aspects through multi-scale supervision and multi-view feature learning in his paper. The former tries to acquire semantic independent features and more generalizable characteristics by using boundary artifacts and noise distribution surrounding tampered regions. The latter enables us to learn from authentic images that are too complicated for the existing semantic segmentation network-based techniques to handle. The model performs relatively better in cross-dataset scenarios and is robust to various post-processing procedures. Using ground truths of forged boundaries and regions, a multi-task fully convolutional network (MFCN) was trained for forgery localization [8]. A hybrid LSTM and ED network model was recently developed in [9] for pixel-wise forgery localization. This method uses spatial characteristics and resampling to capture irregular transitions between fake and real patches. The CNN methods are better at understanding the visual context or learning the semantic features to perform better at image classification tasks. Image manipulation is semantic-agnostic, so such methods tend to be less effective against image manipulation. Also, CNNs have inductive and locality biases and do not capture the global correlation of features. Objectformer [19] and Transforensics [20] are transformer-based methods proposed over the years. These two models' general architecture and design philosophies hardly align with ViT. Specifically, in contrast to ViT, which embeds the patched images immediately before encoding, both approaches extract feature maps using many CNN layers and then use transformers for additional encoding, neglecting important first-hand low-level information.

To improve the existing models' limitations, we designed a dual-branch model that uses handcrafted features, noise and CNN-based features. The dual-branch has RGB info at one branch and noise information fetched by the Bayar convolution on the other, and these features are strengthened and refined by the channel, and spatial attention and ED architecture are further employed to localize manipulated features.

3 Proposed model

The proposed model consists of two parallel branches; one uses an RGB image as an input, while the other uses noise/residual features as an input feature determined by the Bayar convolution and SRM convolution filters. These kernel filters suppress semantic content and enhance low-level manipulation traces (Fig. 2). The channel refines RGB and noise feature representation of manipulation and spatial attention maps generated by their respective modules. The input features are multiplied with the attention maps to strengthen discriminative features. ED architecture is employed where the encoder translates the intermediate features into discriminative feature maps, which are then further processed and recovered by the decoder to generate classification predictions down to the pixel level. Skip connections improve the propagation capabilities of encoder-decoder network features. A dual attention network further improves feature representation by adapting global dependencies along the spatial and channel axes and local semantic characteristics, resulting in a more precise manipulation of localized features. The resulting features from both branches are concatenated for a more accurate final manipulation prediction.

Three components are used in the above model: Residual/Noise filters, ED with skip connection, and visual Attention modules.

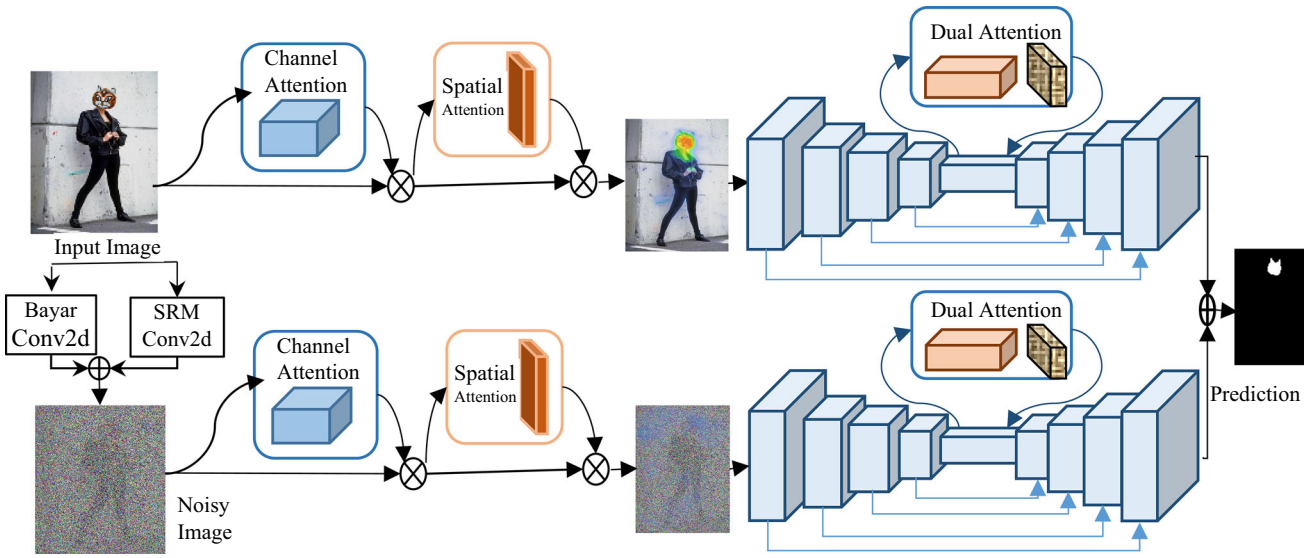


Fig. 2 Overview of the proposed model consisting of a dual branch consisting of RGB and noise branch followed by ED architecture

3.1 Noise inconsistencies

An actual image has uniform noise distribution throughout the image. The intuition behind noise residual as a feature is that when an object is removed from one image/section (of an image) and copied/pasted onto another/section of an image, the noise features of the two are less likely to match. In this configuration, the noise residual is the disparity between a pixel's actual value and its estimated value, which is derived via interpolating the values of nearby pixels, which serves as the model for noise. Bayer and SRM filters are the most standard filters that successfully capture the low-level noise residual features.

3.1.1 Bayar convolution or constrained CNN

Using data, constrained CNN can learn the modifications brought up by image manipulation methods into local pixel relationships. Hence, this approach can suppress image-level content and subsequently learn the latent traces of image manipulation [21]. Constrained CNN's primary purpose is to learn prediction error filters, producing feature maps utilized as low-level forensic traces since they offer superior robustness and universality. To force the CNN to learn the low-level traces, the following constraints are enforced on the weights of the kernel of CNN:

$$\begin{cases} \omega_k^{(l)}(0,0) = -1, \\ \sum\limits_{m,n\neq 0} \omega_k^{(l)}(m,n) = 1, \end{cases} \tag{1}$$

Equation (1) represents the constraints enforced on the filter of the kernel. The superscript indicates the CNN layer. l^{th}, the k^{th} convolutional filter within a layer is indicated by the subscript k, and for a CNN filter, the central value is represented by the spatial index (0,0).

3.1.2 Steganalysis features

Another method used to extract features from an image's noise residuals is SRM (Spatial Rich Models) filters. Fridrich et al. [13] first introduced the concept of SRM. It was primarily created for steganalysis, extracting latent or hidden features from an image's noisy residuals by applying a set number of high-pass filters. Subsequently, those features are combined and sent to ensemble classifiers. It is a specially designed method that essentially calculates the statistics required to extract specific characteristics from the noise residuals surrounding the neighborhood of pixels in an image. This approach yields a feature that can be considered a local noise descriptor.

3.2 Encoder-decoder with skip connection

The architecture used by most semantic segmentation algorithms today is the ED structure. Our encoder network consists of a Vgg-16 network with 13 convolutional layers (w/o fully connected layers), followed by max-pooling layers and is divided into five stages wherein, at each stage, the spatial resolution is halved at every stage, and the channel dimension keeps doubling. This approach allows the model to acquire complex hierarchical representations of visual characteristics, resulting in more reliable and precise predictions. The rationale behind using VGG16(w/o FC layers) as an encoder network is that it has fewer parameters that enable the powerful representation of discriminative visual

features. Hence, an encoder module gathers more semantic information while decreasing the feature mappings and increasing channel dimensions. The decoder network enables up-sampling by mapping low-resolution encoded feature maps to full-scale input-resolution feature maps. The proposed architecture uses the transpose convolution layers for the decoder module to enhance the coarse feature mapping of the full-resolution segmentation map. ED employs skip connection to enhance the neural network feature propagation abilities and to prevent gradient vanishing and exploding gradients [22].

3.3 Attention mechanism

The attention mechanism uses input-dependent weights to replace the traditional learnable fixed weights, allowing the CNN to learn input-aware relationships that help it emphasize the critical features. Two attention modules are used in the model to encode discriminative features more effectively, which are as follows:

3.3.1 Channel attention

A channel attention map is created by utilizing the inter-channel relationship of features. For manipulated localized features, channel attention emphasizes "what" is meaningful for the end task [23]. The module first models the spatial features of various feature maps to generate context descriptors using average and max pooling operations. Subsequently, both descriptors are transmitted to a Multi-layer perceptron (MLP) [23]. Next, the output feature vectors are combined using element-wise summation, and the resulting vector is then normalized using the sigmoid function.

$$\boldsymbol{M_c}(\boldsymbol{F}) = Sigmoid(MLP(AvgPool(\boldsymbol{F})) + MLP(MaxPool(\boldsymbol{F}))) \tag{2}$$

where $\boldsymbol{M_c}(\boldsymbol{F})$ Denotes the feature map that captures channel attention.

3.3.2 Spatial attention

The inter-spatial interaction among features generates the spatial attention map. It is complementary to channel attention and differs in that it concentrates on "where" manipulation localization is presented. Initially, the average-pooling and max-pooling operations are performed on the channel axis to calculate the spatial attention. Subsequently, these results are combined to provide a suitable feature descriptor [23]. The concatenated feature descriptor is subjected to a convolution layer to generate a spatial attention map delineating regions where desired features are prioritized over undesirable ones.

Below are the equations of the spatial attention map.

$$\boldsymbol{M_c}(\boldsymbol{F}) = Sig\big(f^{7\times7}([AvgPool(\boldsymbol{F})); MLP(MaxPool(\boldsymbol{F})])\big) \tag{3}$$

3.3.3 Dual attention mechanism

The Dual Attention network incorporates a self-attention mechanism identifying spatial and channel feature dependencies.

3.3.3.1 Position attention module The purpose of the self-attention mechanism in the position attention module is to capture the spatial relationships between any two positions in the feature map [24]. A local feature $\alpha \in \mathcal{R}^{C\times H\times W}$ applied into the layer of convolution, resulting in the generation of two additional feature maps., $\beta \in \mathcal{R}^{C\times H\times W}$ and $\gamma \in \mathcal{R}^{C\times H\times W}$, respectively. Next, reshape them $\mathcal{R}^{C\times N}$ where $N = H \times W$ denotes the number of pixels of a layer. The spatial attention map $\check{S} \in \mathcal{R}^{N\times N}$ which is then computed by performing a matrix multiplication between the transpose of β and γ and applying a softmax layer:

$$s_{ji} = Sigmoid(\beta_i, \gamma_i)\, where\, i = 1,2, \ldots, N \tag{4}$$

where s_{ji} quantifies the impact of i^{th} position is on j^{th} position. Subsequently, passes the feature α through convolution layers to generate feature $\delta \in \mathcal{R}^{C\times H\times W}$ e and reshape it to a size of $\mathcal{R}^{C\times N}$. Next, calculate the matrix multiplication of δ and the transpose of $\mathbb{S}$ and reshape the resulting matrix to have dimensions$\mathcal{R}^{C\times H\times W}$. Eventually, we determine the product of the input by a scaling factor ρ and aggregate it with the characteristics α using an element-wise summation to obtain the final result. $\theta \in \mathcal{R}^{C\times H\times W}$ Moreover, it is described below:

$$\theta_j = \rho \sum_{i=1}^{N} \big(s_{ji}\delta_i\big) + \alpha_i, \tag{5}$$

The outcome characteristic θ is obtained by taking a weighted sum of the initial characteristics and all the features at each point. Consequently, it selectively combines different contexts based on the spatial attention map and provides a global contextual view.

3.3.3.2 Channel attention module By considering the connections between channel maps, it is feasible to emphasize the interdependence of feature maps and improve the representation of features in a specific semantic domain. Consequently, the channel attention module directly represents channel interdependencies [24]. As opposed to the position attention module, the channel attention map $\acute{Z} \in \mathcal{R}^{C\times H\times W}$

directly from the original features $\mu \in \mathcal{R}^{C\times H\times W}$. To be more precise, we reshape μ to $\mathcal{R}^{C\times N}$ and then multiply μ by its transposition in a matrix. Eventually, a softmax layer is employed to produce the channel attention map $\acute{Z} \in \mathcal{R}^{C\times c}$:

$$z_{ji} = Sigmoid\left(\mu_i, \mu_j\right) where\, i = 1,2, \ldots, C \tag{6}$$

where z_{ji} measures the influence of the i^{th} position's influence on j^{th} position. Furthermore, we apply a matrix multiplication using the transpose of $\acute{Z}$ and μ, transforming the outcome into $\mu \in \mathcal{R}^{C\times H\times W}$. The final output, $F \in \mathcal{R}^{C\times H\times W}$, is then obtained by multiplying the result by a scaling parameter ϑ and using an element-wise sum operation with μ.

$$F_j = \vartheta \sum_{i=1}^{C}\left(x_{ji}\mu_i\right) + \mu_i \tag{7}$$

Equation (7) demonstrates the final feature of each channel, which represents the long-range semantic linkages between feature maps and is computed as a weighted sum of the features from all channels and the original features. It improves the ability to discriminate between different features [24].

3.4 Loss function

Two famous classification and segmentation loss functions have been used: Binary Cross Entropy(BCE) Loss and Focal Loss(FL).

*BCE loss a*ttempts to measure the difference in probability distributions of the estimated and ground-truth masks [25]. It is defined as:

$$\mathcal{L}_{BCE}\left(G_m, \hat{P}_m\right) = -\left(G_m \log\left(\hat{P}_m\right) + (1 - G_m)\log\left(1 - \hat{P}_m\right)\right) \tag{8}$$

where G_m Is the ground truth pixel value of the mask and $\widehat{P}_m$ is the estimated pixel value of the mask. It computes the Loss for every pixel separately and treats every pixel equally.

Focal Loss is another prevalent loss in segmentation tasks and is utilized as an extension of BCE in severely class-imbalanced detection tasks [26]. This allows the model to learn more complicated cases by down-weighting the contribution of less complex ones. Focal Loss uses a modulating factor, p, to down-weight simple examples and concentrate training on hard negatives.

$$\mathcal{FL}_{\gamma,\alpha} = -\left(\alpha(1 - p_i)^{\gamma} y_i log p_i + (1 - \alpha) p_i^{\gamma}(1 - y_i)\log(1 - p_i)\right) \tag{9}$$

where p_i is the predicted probability distribution for the class label, which is unity. An equilibrium factor is α, and parameter γ specifies the rate at which lesser complex data samples are given lesser significance.

The combined loss function is a powerful combination of BCE and focal Loss. BCE loss penalizes the normal distribution of data samples, while focal Loss addresses the problem of class imbalance and focuses on the highly complex task by down-weighting the contribution of easier ones using their hyper-parameter settings.

$$\mathcal{L}_{combined} = \mathcal{L}_{BCE} + \mathcal{FL}_{\gamma,\alpha} \tag{10}$$

Equation (10) represents the combined loss function used for this model. The model uses this integrated loss function during the entire training process.

4 Experiments

This section aims to evaluate the performance of the proposed methodology by validating it on various benchmark datasets and comparing it with various SoTA manipulation localization methods.

4.1 Datasets

4.1.1 Shallowfake dataset

A massive data hunger characterizes deep learning network training. There are insufficient images in the usual datasets used for image manipulation detection today to support deep neural network training. Furthermore, a standard dataset's altered images may not be sufficient for training because they have fewer artifacts. The model undergoes pre-training using CASIAv2, subsequently fine-tuning with additional datasets, and testing is done on them. Table 1 contains the details of the split of the training–testing dataset with their kind of manipulation.

4.1.2 Deepfake dataset

No deepfake image dataset currently contains a ground truth mask for the manipulated regions. Zhang et al. [31] have built their dataset of Faceforensics + + [32], the only deepfake dataset containing masks for most of its videos. Famous FF + + (c23 compressed version used for our work) contains 1000 videos and 5000 fake videos manipulated methods (Deepfakes, Face2Face, Faceshifter, Face-Swap and NeuralTextures). Four manipulations are considered for the frames extracted as 1000 face shifter videos contain no ground truth mask from the videos. Two frames are extracted for each video, and due to some accessibility concerns prevented us

Table 1 Training and testing split of the various benchmark datasets. S means splicing, C means copy-move, and R means removal

Datasets	Training Set	Testing Set	Total Samples	Types of manipulation
CASIA V2.0 [27]	5063	–	5063	S, C
CASIA V1.0 [27]	–	920	920	S, C
COLUMBIA [28]	130	50	180	S
COVERAGE [29]	75	25	100	C
NIST16 [30]	414	150	564	S, C, R

from downloading some legitimate and bogus videos. We have obtained a total of 8449 genuine frames and 7330 counterfeit frames.

4.2 Experimental setup

We use the Vgg-Net16 [32] as a backbone for encoder-decoder architecture, pre-trained on ImageNet [33], to develop our model using the PyTorch framework. The model is executed using two NVIDIA RTX A5000 GPUs, and the image is scaled to 256 × 256. The model is optimized using the Adam optimizer with a batch size 32 during the training and testing phase. The starting learning rate is 1e-4, which decays every 10th step with a decay rate of 0.8.

4.2.1 Quantitative analysis

We have quantitatively assessed the performance of the shallowfake and deepfake datasets.

4.2.1.1 Shallowfake dataset Following the approach [34], the model was trained using the CASIA2 dataset and fine-tuned with standard shallowfakes datasets like Nist16, Coverage, Columbia, and CASIA1. The AUC and F_1 scores of these datasets are recorded in Table 2. For comparison, two categories of models are considered, i.e. unsupervised and DNN models. The table shows that the model outperforms the unsupervised models by a significant margin. The performance of handcrafted features intended for the specific type of manipulation is severely constrained, and all of these traditional methods capture certain tampering artifacts with limited information for detection. Our method's scores are comparable and outperform the other DNN-based methods at various datasets. For the CASIA and Columbia, our model performs significantly well with an AUC score even reaching 96.72%, and on the Nist16 dataset, it has a comparable score to other models (AUC ~ 99.9%).

In comparison, the model performs poorly on the COVERAGE dataset, with an AUC score reaching 77.38% owing to fewer images and the dataset containing copied/moved objects of similar appearance. Our technique captures the broader range of features like RGB features, noise inconsistencies, and global context rather than the neighboring pixels, which helps gather more information for manipulation classification. Various complex CNN models cannot perform well; this could be attributed to most DNN-based methods that model the network by superimposing multiple CNN networks or adding complex branches. For instance, TDA-Net combines three CNN streams, such as end-to-end training for a complex network, which makes it harder for the network to train and requires more computing power. Also, few models are so much simpler that they focus only on the semantic features and hence fail to locate the tampered segments accurately. On the other hand, our model is relatively less complex, easily captures non-semantic features and does not require significant training data to achieve comparable performance.

4.2.2 Deepfake dataset

Ten models are considered for evaluation on the deepfake dataset. Six are the SoTA image manipulation models; the others are the standard image-segmentation models. Codes of the models that are available on GitHub are considered for comparison. Pre-trained Pytorch models are used for image segmentation models, which are further fine-tuned. Three evaluation metrics are used for a more comprehensive evaluation. Table 3 shows the experimental results of the deepfake on the various models. All the models performed decently except MantraNet [10] on the various categories of the Face-forensics + + dataset. It could be attributed to the fact that most of the manipulations are of the face, and there is a single entity that occupies the entire frame, and all the models are powerful enough to capture such apparent artifacts. MantraNet performs poorly on all categories of the deepfake, possibly because images/frames are not high-resolution and contain blurriness and noise, which the model cannot process and learn the latent discriminative features. NedB-Net [16] is another model that has performed well but has a relatively lower score than other SoTA models, which may be due to the low-quality images and the more significant manipulated regions; the paper's authors also highlight this problem. DL-Net [44] performs reasonably well in the FF + + dataset, with the F1 score reaching 96%, owing to its capability to capture high and low-level cues using noise level segmentation map prediction, which constraints the model to focus on

Table 2 Evaluation experiments results of various models on the shallowfake dataset

Category	Method	NIST16		COLUMBIA		COVERAGE		CASIA	
		AUC	F1	AUC	F1	AUC	F1	AUC	F1
Unsupervised	ELA [35]	0.429	0.236	0.581	0.470	0.583	0.222	0.613	0.214
	NOI1 [19]	0.487	0.285	0.546	0.574	0.587	0.269	0.612	0.263
	CFA1 [36]	0.501	0.174	0.720	0.467	0.485	0.190	0.522	0.207
DNN-based model	MFCN [8]	–	0.571	–	0.612	–	–	–	0.541
	RGB-N [37]	0.937	0.722	0.858	0.697	0.817	0.474	0.795	0.582
	J-LSTM [38]	0.764	–	–		0.712	–	–	–
	LSTM-EnDec [9]	0.794	–		–	0.712	–	–	–
	CR-CNN [39]	0.992	0.927	0.861	0.790	0.939	0.757	0.789	0.475
	ManTra-Net [10]	0.795	–	0.824	–	0.819	–	0.817	–
	TDA-Net [34]	0.948	0.756	0.892	0.735	0.864	0.474	0.831	0.582
	GSR-Net [40]	0.945	0.736			0.768	0.489	0.796	0.574
	SPAN [41]	0.961	0.582	0.936	0.815	0.937	0.558	0.838	0.382
	PSCC-Net [42]	0.996	0.819	–	–	0.941	0.723	0.875	0.554
	MVSS-Net + + [18]	0.976	0.854	–	–	0.897	0.753	0.844	0.546
	ObjectFormer [19]	0.996	0.824	–	–	0.957	0.7580	0.882	0.579
	TA-Net [43]	**0.997**	0.865	–	–	**0.978**	**0.782**	0.893	0.614
	Transforensics [20]	–	–	–	–	0.884	0.674	0.850	0.627
	Our model	0.9919	**0.9910**	**0.9672**	**0.9556**	0.7738	0.4362	**0.9494**	**0.8765**

"–" means unknown score
Bold values represent the highest score among the compared models

the manipulated regions. However, the model has a low score for Face-swap manipulation among the four categories of FF + + manipulation. Another method [45] used for deepfake localization uses a weak supervision framework and uses three methods, i.e. GradCAM, Patches and Attention, for results illustrations. We have used GradCAM methods for score comparison. The method performs outstanding well in the weak supervision setting, showing the powerful discriminative capabilities of the model. However, like the earlier model, the model has a drop in performance for the FS category of manipulation, which could be the network's inability to make accurate predictions at the edges. DADF method [46] performs better than most models, which use multi-scale adapters to capture short and long-range forgeries and guided attention mechanisms, enhancing rich forgery clues. Their scores are at par with other methods and a proper State-of-the-art method for comparison. Another method [31] used also performed well, with the F1 score reaching 98%. Their method is built on top of existing UperNet and uses Bayar convolution methods to trace noise clues. Despite being these State-of-the-art models, all models score more than 90%, which could be attributed to the fact that most of the manipulation has been done on the face, which is easy for the models to recognize. Our model has performed considerably well in the various categories of the FF + + dataset and outperformed the scores of other standard models. These evaluations showed that different modules designed well to work in tandem resulted in the robust learning of the discriminative for varied manipulation.

4.3 Qualitative analysis

This section compares our method with the two most competitive methods, i.e. MantraNet and MVSSS, to give some specific qualitative outcomes on both shallowfakes and deepfake datasets. Figure 3 shows the visualization results of image manipulation detection. Our method achieves better localization results than other methods, as the other methods generate many false positives. In the case of the shallowfake dataset, the method has superior localization performance for CASIA, Columbia and Nist16 datasets. For the deepfake dataset, localization seems much easier, as most image manipulation is done on the face, making it easy for localization. MantraNet exhibits a significant departure from the ground truth, whereas MVSSNET exhibits a significant rate of false positives in regions that have not been altered. The primary cause of this outcome is that during the training phase, MVSSNET used a lot of natural images, which could have a clear negative impact on network training. Additionally, MVSSNET dramatically increases false positives due to

Table 3 Evaluation experiments results of various models on categories of F + + dataset. IMD means Image Manipulation Detection models, and IS means Image Segmentation Models

Types	Methods	Deepfakes			Face2face			Faceswap			Neural texture		
		IoU	AUC	F1	IoU	AUC	F1	IoU	AUC	F1	IoU	AUC	F1
IMD models	MVSS-Net [47]	0.9558	0.9996	0.9787	0.9788	0.9997	0.9893	0.9570	0.9989	0.9780	0.9382	0.998	0.968
	MantraNet [10]	0.3469	0.9523	0.5151	0.3553	0.9159	0.5243	0.3394	0.8888	0.5068	0.3664	0.9706	0.5363
	NedB_Net [16]	0.8811	0.9767	0.9368	0.8432	0.9662	0.9149	0.8522	0.9700	0.9221	0.8827	0.9783	0.9377
	DL-Net [44]	0.8750	0.9952	0.9337	0.9108	0.9946	0.9533	0.8976	0.9976	0.9460	0.9262	0.9979	0.9617
	Weakly_Super_Gradcam [45]	0.9787	0.9990	0.9897	0.8753	0.9795	0.9336	0.9575	0.9791	0.9231	0.9861	0.9990	0.9868
	DADF [46]	0.9453	0.9939	0.9677	0.9621	0.9899	0.9786	0.9599	0.9896	0.9655	0.9236	0.9788	0.9586
	ShallowDeepfake_local [31]	0.9617	0.999	0.9713	0.9801	0.9898	0.9799	0.9485	0.9856	0.9666	0.9365	0.9989	0.9689
IS Models	DeepLab [48]	0.9428	0.9981	0.9706	**0.9840**	**0.9999**	**0.9919**	0.9769	0.9986	0.9883	0.9704	0.9998	0.9640
	FCN [49]	0.9701	0.9967	0.9848	0.9834	0.9988	0.9786	0.9470	0.9991	0.9728	0.9591	0.9984	0.9728
	LRASPP [50]	0.9114	0.9992	0.9536	0.9445	0.9998	0.9714	0.9106	0.9995	0.9532	0.9399	0.9997	0.9690
Our model		**0.9736**	**0.9940**	**0.9866**	0.9820	0.9964	0.9909	**0.9810**	**0.9983**	**0.9904**	**0.9710**	**0.9913**	**0.9853**

Bold values represent the highest score among the compared models

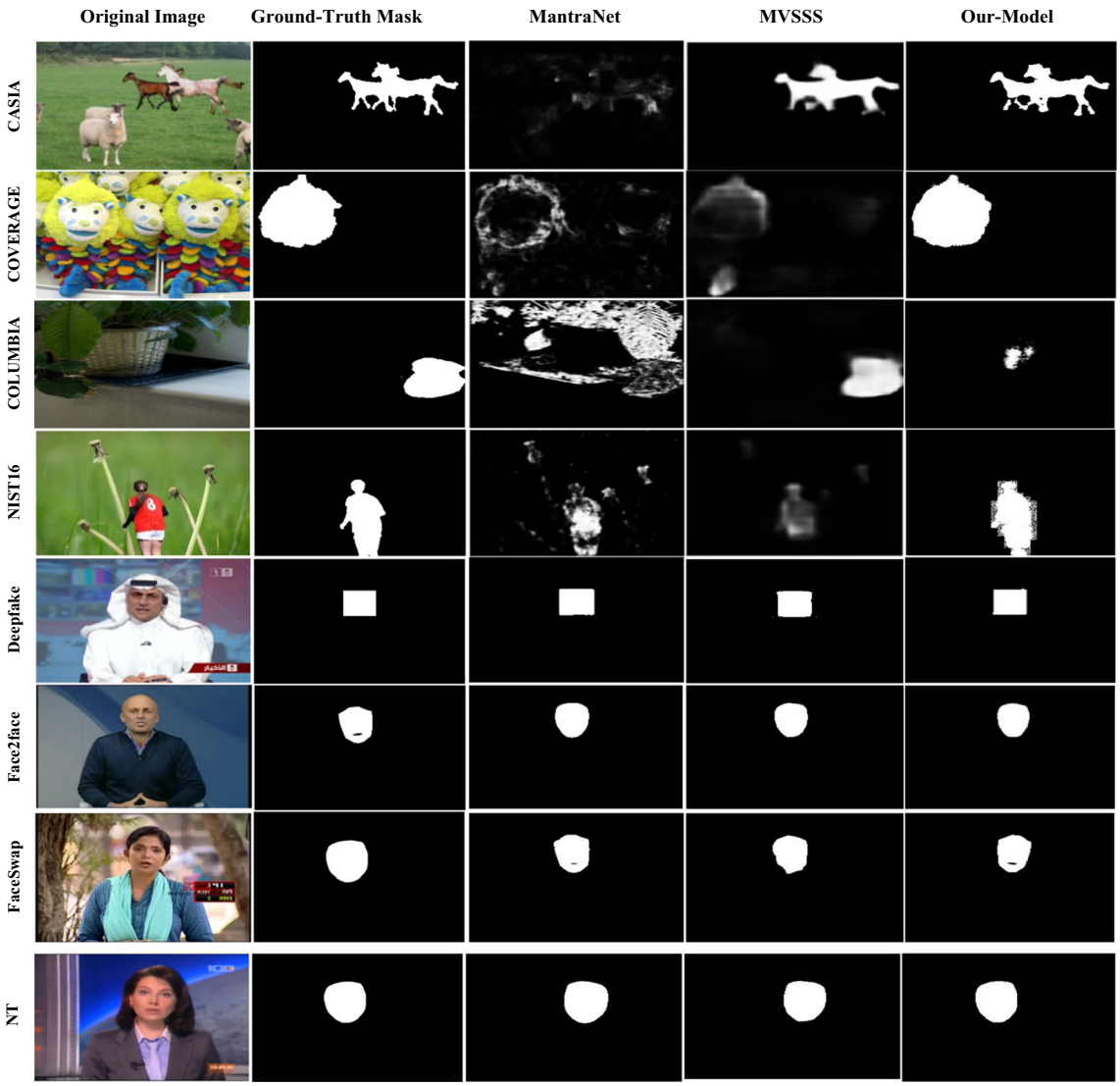


Fig. 3 Qualitative visualization results of the manipulation localization for shallow fakes and deepfake dataset images for different methods

frequently responding to altered and non-manipulated portions of the image. On the contrary, our method focuses on low-level features like noise and high-level contextual features, leading to better localization of manipulated artifacts.

4.4 Ablation studies

We assess the proposed network in various settings with the components introduced one at a time to study the impact of each component. All the components were trained on the CASIA2 dataset and then evaluated on the different shallowfakes datasets, i.e. NIST16, Columbia, Coverage and CASIA1. Table 4 shows the results of the ablations experiment. The different experimental settings are discussed below:

Case A: Model without Channel Attention Channel attention modules have been removed from both branches to study their impact. AUC and F_1 scores decrease slightly, showing inter-relation along the channel axis containing potential features that assisted in the detection task. Two things worth noting: for the Columbia and Nist16 dataset, scores decreased very slightly, nearly ~ 2%(AUC score), while for the Coverage dataset, it decreased significantly by nearly ~ 9%, showing that where data samples are less, attention mechanism plays a pivotal role.

Case B: Model without Spatial Attention From scores, it perceives that spatial attention contributes less in comparison to channel attention to the detection features. It could be attributed to spatial attention focusing on high-level or semantic features, which is less critical for manipulation.

Table 4 Ablation experiment of different components where the model trained on CASIA2 and tested on other datasets

Cases	NIST16		COLUMBIA		COVERAGE		CASIA	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Case A	0.9246	0.9688	0.9516	0.8758	0.6891	0.3623	0.8868	0.8078
Case B	0.9327	0.9896	0.9626	0.9285	0.7208	0.3953	0.9123	0.852
Case C	0.8055	0.8302	0.8328	0.7497	0.7373	0.3946	0.8033	0.4254
Case D	0.9087	0.9291	0.93	0.8398	0.7307	0.4049	0.9210	0.8560
Case E	0.8688	0.8015	0.8873	0.7333	0.7261	0.3871	0.7543	0.4852
Case F	0.9239	0.9488	0.9435	0.8569	0.6052	0.2879	0.8863	0.8562
Overall model	0.9441	0.9952	0.9641	0.8988	0.7542	0.4272	0.9313	0.8650

Table 5 Computation complexity analysis of different models

Methods	Parameters (millions)	CPU time (sec)	GPU time (sec)
MantraNet	3.80	8.06	0.14
MVSS-Net	142.782	9.66	1.13
SPAN	4.06	8.08	0.278
Nedb-Net	45.085	10.11	1.171
DeepLabv3_ResNet50	39.63	9.97	1.17
FCN_ResNet50	32.946	10.20	1.14
LRASPP_MobileNet	3.21	9.01	1.01
Our_model	38.86	9.09	0.94

GPU and CPU time are measured for an epoch of batch size of 32

Also, one of the trivial observations is that scores declined slightly and uniformly across all the datasets.

Case C: Model without Channel and Spatial Attention In this case, channel and spatial have been removed from both branches to study their relevance. Scores have decreased significantly, nearly ~ 14%, across all the datasets except the Coverage dataset. It shows that these modules work better together to capture informative features and global correlations along every axis and emphasize such critical features.

Case D: Model without Dual Attention Dual attention modules are axed from the Encoder-decoder architecture to investigate their importance. This dual attention contributes less to the overall detection task of either the spatial or channel attention module. Also, the dual attention works almost uniformly for every dataset, showing its ability to capture long-range contextual information for varied manipulations.

Case E: Model without Noise branch Here, the noise branch, consisting of the RGB branch, is completely removed from the model. This means that RGB high-level or semantic features are primarily used for detection. Here, the performance decreased drastically by nearly ~ 8%, confirming that low-level features like noise consistencies are crucial indicators in the overall detection task. For the CASIA model, scores have decreased to a greater extent, around ~ 19%, as it varied manipulation like splicing and copy-move, showing the relevance of residual noise feature to varied manipulation.

Case F: Model without RGB branch The RGB branch is completely removed from the overall model, and the noise branch is used for manipulation detection. In this scenario, the AUC score is decreased to a smaller extent, around 4%, again confirming that the low-level noise features have more relevance than high-level semantic features for the detection task. For the Coverage dataset, the performance has decreased drastically, meaning that noise branches need ample datasets to learn noise inconsistencies for detection tasks.

5 Computational complexity analysis

In this section, we assessed the complexity of three distinct deep networks: MantraNet, SPAN, MVSSNET, Nedb_Net [16], DeepLabv3_ResNet50 [48], FCN_ResNet50 [51] and LRASPP_MobileNet [50]. All experimental analysis is done on two NVIDIA RTX A5000 GPUs. Table 5 shows the computational analysis of different models compared to ours. Regarding the number of parameters, the model is relatively lighter, with 38.86 million parameters compared to MVSS-Net and Deeplab. MantraNet and LRASPP are many relatively lighter models than other models. Another metric was calculated in CPU and GPU time for one epoch for a batch size of 32 images of the CASIA dataset. The model

runs on the GPU and CPU to measure their time in seconds. GPU time is always less time for different models. Our model takes less time both on GPU and CPU.

6 Conclusion

This paper proposes a novel dual-branch architecture consisting of Noise Residual extraction modules at one branch and RGB information at the other branch powered by the attention mechanism before feeding to the ED architecture for precise IML prediction. The model effectively captures the low-level inconsistencies critical for IML tasks and additional semantic features. Extensive experiments on the shallowfakes and deepfake datasets have shown that the model is able to capture subtle traces of manipulation and achieves SoTA results. Future work may involve checking the model's generalizability on the unseen dataset and robust evaluation of various compression scenarios.

Author contributions D.D.: Software, Validation, Investigation, Data Curation, Writing – Original Draft, Visualization. D.K.V.: Conceptualization, Methodology, Formal Analysis, Resources, Writing – Review & Editing, Supervision, Project Administration, Funding Acquisition.

Funding No funding was received for this research.

Data availability Data is open-source, and no prior permission is required. It is available through these links. CASIA: casia dataset (kaggle.com) Nist16: OpenMFC (nist.gov) Coverage: wenbihan/coverage: Copy-Move forgery database with similar but Genuine objects. ICIP2016 paper (github.com) Columbia: Columbia Image Splicing Detection Evaluation Dataset Faceforensics + + : ondyari/FaceForensics: Github of the FaceForensics dataset.

Declarations

Conflict of interests The authors declare no competing interests.

Ethical approval Not Applicable.

References

1. J. Zhang, H. Tohidypour, Y. Wang and P. Nasiopoulos, "Shallow- and Deep- fake Image Manipulation Localization Using Deep Learning," In *International Conference on Computing, Networking and Communications (ICNC)*, Honolulu, HI, USA, 2023.
2. S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," In *3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, Malaysia, 2015.
3. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," In *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009.
4. Li, S., Xu, S., Ma, W., Zong, Q.: Image manipulation localization using attentional cross-domain CNN features. IEEE Trans. Neural Netw. Learning Syst. **34**(9), 5614–5628 (2023)
5. Y. Wu, W. AbdAlmageed and P. Natarajan, "ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features," In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.
6. Z. Zhang, Y. Qian, Y. Zhao, X. Zhang, L. Zhu, J. Wang and J. Zhao, "Noise and Edge Based Dual Branch Image Manipulation Detection," In *International Conference on Computing, Networks and Internet of Things*, New York, 2023.
7. Kong, C., Chen, B., Li, H., Wang, S., Rocha, A., Kwong, S.: Detect and locate: exposing face manipulation by semantic—and noise-level telltales. IEEE Trans. Inf. Forensics Secur. **17**, 1741–1756 (2022)
8. D. Tantaru, E. Oneata and D. Oneata, "Weakly-supervised deepfake localization in diffusion-generated images," In *IEEE Workshop on Applications of Computer Vision (WACV)*, Hawai, 2024.
9. Lai, Y., Luo, Z., Yu, Z.: Detect Any Deepfakes: Segment Anything Meets Face Forgery Detection and Localization. In: Kang, W., Pan, Z., Ben, X., Bian, Z., Shiqi, Y., He, Z., Wang, J. (eds.) Biometric Recognition. Springer, Singapore (2023)
10. L.-C. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," in *arxiv.org/abs/1706.05587*, 2017.
11. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2017)
12. A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam and Q. Le, "Searching for MobileNetV3," In *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019.
13. J. Dong, W. Wang and T. Tan, "CASIA Image Tampering Detection Evaluation Database," In *IEEE China Summit and International Conference on Signal and Information Processing*, Beijing, China, 2013.
14. T.-T. Ng, J. Hsu and S.-F. Chang, "Columbia Image Splicing Detection Evaluation Dataset," In *DVMM Laboratory of Columbia University*, 2009.
15. B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen and S. Winkler, "COVERAGE—A novel database for copy-move forgery detection," In *IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, USA, 2016.
16. H. Guan, M. Kozak, E. Robertson, Y. Lee, A. N. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith and J. Fiscus, "MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation," In *IEEE Winter Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI, USA, 2019.
17. Neal Krawetz and Hacker Factor Solutions: A picture's worth digital image analysis and forensics. Hacker Factor Solut. **2**(2), 2 (2007)
18. Mahdian, B., Saic, S.: Using noise inconsistencies for blind image forensics. Image Vis. Comput. **27**(10), 1497–1503 (2009)
19. Ferrara, P., Bianchi, T., Rosa, A.D., Piva, A.: Image forgery localization via fine-grained analysis of CFA artifacts. IEEE Trans. Inf. Forensics Secur. **7**(5), 1566–1577 (2012)
20. Salloum, R., Ren, Y., Kuo, C.-C.J.: Image splicing localization using a multi-task fully convolutional network (MFCN). J. Vis. Commun. Image Represent. **51**, 201–209 (2018)
21. P. Zhou, X. Han, V. I. Morariu and L. S. Davis, "Learning Rich Features for Image Manipulation Detection," In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.
22. J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj and B. Manjunath, "Exploiting Spatial Structure for Localizing Manipulated Image Regions," In *IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

23. Bappy, J.H., Simons, C., Nataraj, L., Manjunath, B.S., Roy-Chowdhury, A.K.: Hybrid LSTM and encoder–decoder architecture for detection of image forgeries. IEEE Trans. Image Process. **28**(7), 3286–3300 (2019)
24. C. Yang, H. Li, F. Lin, B. Jiang and H. Zhao, "Constrained R-Cnn: A General Image Manipulation Detection Model," In *IEEE International Conference on Multimedia and Expo (ICME)*, London, UK, 2020.
25. P. Zhou, B.-C. Chen, X. Han, M. Najibi, A. Shrivastava, S.-N. Lim and L. Davis, "Generate, Segment, and Refine: Towards Generic Manipulation Segmentation," In *Proceedings of the AAAI Conference on Artificial Intelligence*, New York, USA, 2020.
26. X. Hu, Z. Zhang, Z. Jiang, S. Chaudhari, Z. Yang and R. Nevatia, "SPAN: Spatial Pyramid Attention Network for Image Manipulation Localization," In *European Conference on Computer Vision*, Glasgow, United Kingdom, 2020.
27. Liu, X., Liu, Y., Chen, J., Liu, X.: PSCC-Net: progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Trans. Circuits Syst. Video Technol. **32**(11), 7505–7517 (2022)
28. Dong, C., Chen, X., Hu, R., Cao, J., Li, X.: MVSS-Net: multi-view multi-scale supervised networks for image manipulation detection. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3539–3553 (2022)
29. J. Wang, Z. Wu, J. Chen, X. Han, A. Shrivastava, S.-N. Lim and Y.-G. Jiang, "ObjectFormer for Image Manipulation Detection and Localization," In *Computer Vision and Pattern Recognition*, New Orleans, 2022.
30. Shi, Z., Chen, H., Zhang, D.: Transformer-auxiliary neural networks for image manipulation localization by operator inductions. IEEE Trans. Circuits Syst. Video Technol. **33**(9), 4907–4920 (2023)
31. J. Hao, Z. Zhang, S. Yang, D. Xie and S. Pu, "TransForensics: Image Forgery Localization with Dense Self-Attention," In *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021.
32. X. Chen, C. Dong, J. Ji, J. Cao and X. Li, "Image Manipulation Detection by Multi-View Multi-Scale Supervision," In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC Canada, 2021.
33. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2016)
34. Dagar, D., Vishwakarma, D.K.: A literature review and perspectives in deepfakes: generation, detection, and applications. Int J Multimed Inf Retr **11**, 219–289 (2022)
35. Verdoliva, L.: Media forensics and deepfakes: an overview. IEEE J Sel Top Signal Process **14**(5), 910–932 (2020)
36. Q. Bammey, R. G. v. Gioi and J.-M. Morel, "An Adaptive Neural Network for Unsupervised Mosaic Consistency Analysis in Image Forensics," In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.
37. Iakovidou, C., Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y.: Content-aware detection of JPEG grid inconsistencies for intuitive image forensics. J. Vis. Commun. Image Represent. **54**, 155–170 (2018)
38. Muhammad, G., Al-Hammadi, M.H., Hussain, M., Bebis, G.: Image forgery detection using steerable pyramid transform and local binary pattern. Mach. Vis. Appl. **25**, 985–995 (2013)
39. T. Carvalho, F. A. Faria, H. Pedrini, R. d. S. Torres and A. Rocha, "Illuminant-Based Transformed Spaces for Image Forensics," *IEEE Transactions on Information Forensics and Security,* vol. 11, no. 4, pp. 720–733, 2016.
40. Y. Rao and J. Ni, "Self-supervised Domain Adaptation for Forgery Localization of JPEG Compressed Images," In *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021.
41. X. Ma, B. Du, Z. Jiang, A. Y. A. Hammadi and J. Zhou, "IML-ViT: Benchmarking Image Manipulation Localization by Vision Transformer," in *arxiv.org/abs/2307.14863*, 2023.
42. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Trans. Inf. Forensics Secur. **7**(3), 868–882 (2012)
43. Z. Zhang, Y. Qian, Y. Zhao, L. Zhu and J. Wang, "Noise and Edge Based Dual Branch Image Manipulation Detection," In *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things*, New York US, 2023.
44. Cozzolino, D., Verdoliva, L.: Noiseprint: a CNN-based camera model fingerprint. IEEE Trans. Inf. Forensics Secur. **15**, 144–159 (2019)
45. Bayar, B., Stamm, M.C.: Constrained convolutional neural networks: a new approach towards general purpose image manipulation detection. IEEE Trans. Inf. Forensics Secur. **13**(11), 2691–2706 (2018)
46. Z. Lai, H. Sun, R. Tian, N. Ding, Z. Wu and Y. Wang, "Rethinking Skip Connections in Encoder-decoder Networks for Monocular Depth Estimation," in https://doi.org/10.48550/arXiv.2208.13441, 2022.
47. S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, "CBAM: Convolutional Block Attention Module," In *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, 2018.
48. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang and H. Lu, "Dual Attention Network for Scene Segmentation," In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.
49. M. Yi-de, L. Qing and Q. Zhi-bai, "Automated image segmentation using improved PCNN model based on cross-entropy," In *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, Hong Kong, China, 2004.
50. T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," In *IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.
51. D. Cozzolino, D. Gragnaniello and L. Verdoliva, "Image forgery detection through residual-based local descriptors and block-matching," In *IEEE International Conference on Image Processing (ICIP)*, Paris, 2014.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Energy Systems
https://doi.org/10.1007/s12667-024-00682-8

ORIGINAL PAPER

Short-term aggregate electric vehicle charging load forecasting in diverse conditions with minimal data using transfer and meta-learning

Shashank Narayana Gowda[1] · Keshav Nath[2] · Chen Zhang[1] · Rohan Gowda[3] · Rajit Gadh[1]

Received: 16 October 2023 / Accepted: 12 April 2024
© The Author(s) 2024

Abstract
The proliferation of electric vehicles (EVs) necessitates accurate EV charging load forecasting for demand-side management and electric-grid planning. Conventional machine learning-based load forecasting methods like long short-term memory (LSTM) neural networks rely on large amounts of historical data, which can be resource-intensive and time-consuming to collect. In this study, we employ Transfer Learning (TL) and Model-Agnostic Meta-Learning (MAML) for short term EV charging load forecasting. These methods involve pre-training a base model on a larger comprehensive EV charging dataset followed by fine-tuning using a few days' worth of EV charging data in our target location. We find that the performance of both the TL and MAML models outperform traditional LSTM models and other classic machine learning models in the context of forecast accuracy when working in three different settings with limited data , load variance, and diverse geographical locations. The error metrics from TL and MAML are up to 24% and 61% lower than deep learning and classic machine learning models respectively.

Keywords Forecasting · Transfer learning · Meta-learning · Machine learning · Electric vehicles · Electric load

1 Introduction

The increasing incorporation of renewable energy sources (RES) and electric vehicles (EVs) introduces complexities in achieving coordination between electricity supply and demand due to their inherent uncertainties. Consequently, accurate forecasting of electricity generation and loads assumes pivotal importance in sustaining a stable smart grid, enabling the precise matching of supply and demand across various time scales [1]. On the supply side, wind and solar energy sources exhibit intermittency, relying on external weather variables outside human control.

Extended author information available on the last page of the article

Published online: 14 June 2024

Simultaneously, the forecasting of EV charging loads becomes a necessity for long-term infrastructure planning and medium-to-short-term demand-side management. These efforts mitigate peak loads, reduce costs, and minimize the curtailment of RES [2]. Furthermore, harnessing demand-side flexibility has the potential to yield cost savings in both planning and operational aspects [3].

Traditionally, EV charging load forecasting is treated as a time series problem and involves the application of statistical and machine learning approaches [2]. Among statistical methods, autoregressive integrated moving average (ARIMA) and exponential smoothing (ETS) are popular due to their simplicity, robustness, efficiency, and accuracy [4]. Machine learning techniques, such as Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN) that leverages both current and stored information in memory, are widely used for time series forecasting, at times surpassing the performance of statistical methods [4]. To enhance the predictive accuracy further, the authors in [2] adopted ensemble learning techniques that combine the strengths of multiple machine learning models, effectively synergizing their capabilities to enhance the precision of EV charging load forecasting.

However, neural network (NN)-based forecasting methods exhibit suboptimal performance in scenarios characterized by limited training data. This limitation often arises when data collection proves challenging, expensive, subject to strict regulations, or when dealing with newly established setups [5]. One approach to address this challenge is transfer learning (TL), which leverages pre-trained forecasting models developed using extensive and diverse datasets as initial models. These pre-trained models can then undergo fine-tuning using a limited dataset specific to the problem at hand [6]. [7] proposed the use of transfer learning using deep generative models for cold-start forecasting of residential EV charging behavior. Generative models were used to augment the small samples in the target dataset before fine-tuning. A hybrid LSTM transfer learning model was used in [8] to predict EV charging demand and network voltage profiles with limited data within a single city.

Another promising algorithm for mitigating the data scarcity issue is model-agnostic meta-learning (MAML). MAML is capable of training on a variety of learning tasks, enabling it to adeptly tackle new learning tasks with a limited number of available training samples [9]. [10] introduced a transferable MAML model tailored for short-term load forecasting in individual Australian households. [5] harnessed transfer and meta-learning techniques to predict personalized residential loads in Korea and substation loads in Portugal but ignored load patterns during weekends and holidays. Both [10] and [5] performed univariate load forecasting where the only feature used to predict load was historical load. [11] employed a meta-learning model to forecast load profiles at the transmission level, utilizing open-access data from Elia, a Belgian transmission system operator. In [11], the forecast was performed to find the future values in the same source dataset, meanwhile [10] and [5] forecasted loads of similar scale in the same country.

In this study, we employ TL and MAML techniques for short-term aggergate EV charging load forecasting. Our investigation spans two distinct geographical areas: the Calfifornia Institute of Technology (CalTech) campus [12] and the Trondheim region in Norway [13]. We accomplish this by leveraging a comprehensive EV charging load dataset sourced from Boulder, Colorado for pre-training [2]. These

datasets encompass varying scales, including campus-level, housing initiative-level, and city-level data. Moreover, they originate from different locations, namely Pasadena, CA, USA; Trondheim, Norway; and Boulder, CO, USA. Additionally, we work with aggregated EV charging data in both shared charger and private charger settings. To assess the effectiveness of our approach, we evaluate the performance of different TL, MAML, and traditional LSTM network variants on these diverse datasets at the time scales of 10–20 days. The models only get to be fine-tuned on 10 or 20 days' worth of data to predict the next 5–10 days on each of the investigated datasets. It is worth noting that, to the best of our knowledge, our study represents the first of its kind in this context. We compare our results to traditional LSTM deep learning networks and to other classic machine learning models.

The key contributions of our paper can be summarized as follows:

- We use TL and MAML training to create short-term load forecasting models;
- Additional temporal features such as seasons, weekends and holidays are incorporated to create richer data;
- Utilize adjacent time windows as query and support sets for MAML;
- Models utilize small amounts (10 or 20 days) of fine-tuning data;
- We ensure the robustness of models in making predictions across varying scales, geographies and demographics by utilizing multiple datasets;

The subsequent sections of this paper are structured as follows:

Methods: This section provides a comprehensive overview of the diverse models employed for short-term EV charging load forecasting. Additionally, it outlines the particulars of the datasets, and delves into the data pre-processing procedures, model architectures, hyperparameters, and details concerning the experimental configuration.

Results: In this section, we present the outcomes of our study and highlight the noteworthy observations derived from the experiments.

Discussion: The concluding section summarizes our findings and provides prospective directions for future research endeavors.

2 Methods

2.1 Datasets

We utilized three distinct publicly available datasets:

2.1.1 Colorado dataset

The dataset from Colorado [14] used in the work of [2], encompasses records of individual EV charging transactions recorded at city-owned shared charging stations located in Boulder, Colorado, USA, covering a time frame extending from January

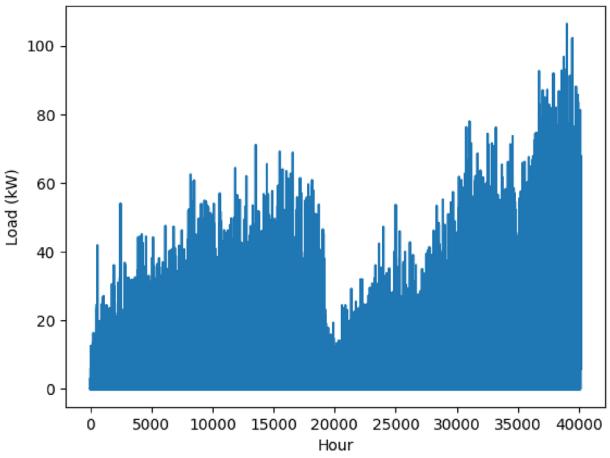


Fig. 1 Aggregate EV charging load distribution in the Colorado dataset

1, 2018, to July 31, 2022. The aggregated EV charging load data is visually represented in Fig. 1.

2.1.2 ACN dataset

The dataset from the Adaptive Charging Network (ACN), as detailed in [12], comprises records of EV charging transactions logged at the CalTech and JPL campuses. Our particular focus centers on the CalTech data, primarily because it exhibits lower EV penetration and a more sparse nature, as visually depicted in Fig. 2. This dataset encompasses a timeframe ranging from April 4, 2019, to September 16, 2021.

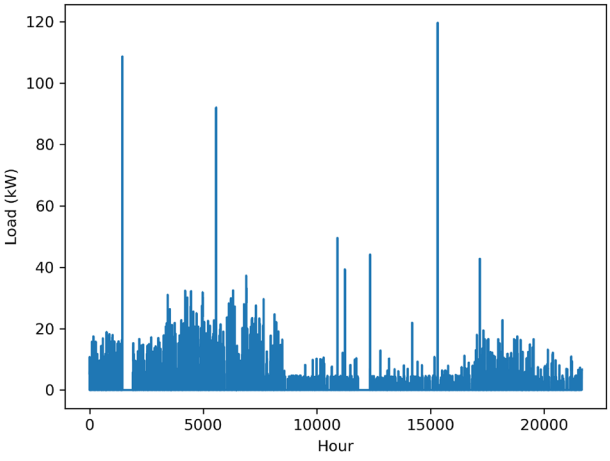


Fig. 2 Aggregate EV charging load distribution in the ACN dataset

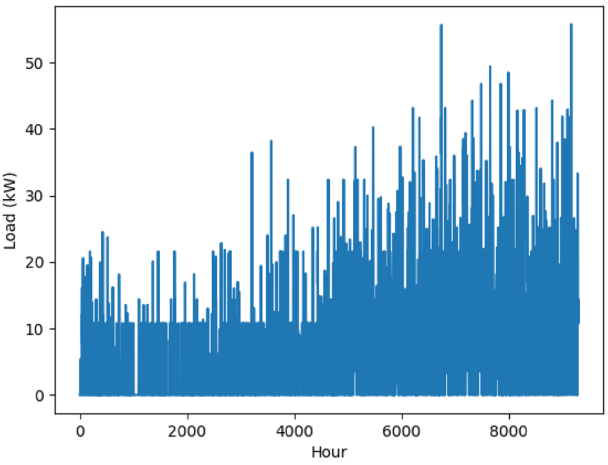


Fig. 3 Aggregate EV charging load distribution from shared chargers in the Norway dataset

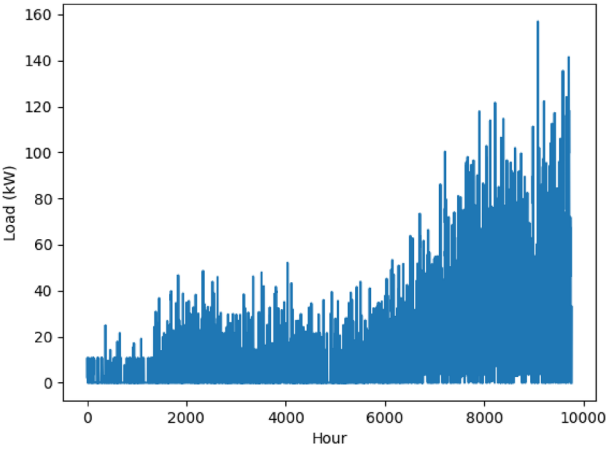


Fig. 4 Aggregate EV charging load distribution from private chargers in the Norway dataset

2.1.3 Norway dataset

The Norway dataset [15] used in [13] contains hourly charging load data from charging stations within a residential block in the Trondheim region of Norway. It consists of two parts—aggregate shared charger data and aggregate private charger data visualized in Figs. 3 and 4 respectively. Both these datasets exhibit an irregular and challenging data distribution because chargers were added in each of these settings over time. Additionally, we can see that the magnitude of the aggregated private charger data is about three times the aggregated shared charger data. The aggregate shared charger data was collected from January 10, 2019 until January

31, 2020 while the aggregate private charger was collected from December 21, 2018 until January 31, 2020.

2.2 Data pre-processing

We transformed both the Colorado and ACN datasets, originally containing charging transactions, into an hourly load format. Following this, all four datasets (Colorado, ACN, Norway-shared, and Norway-private) underwent an extensive pre-processing procedure, incorporating a comprehensive set of 10 features encompassing hour, quarter-of-day, day, day-of-week, week-of-year, month, quarter-of-year, season, and load (kW).

The data was then structured into input–output pairs, with the input comprising a sequence of 24 consecutive hours as features, with the load of the 25th hour serving as the target variable. $D = (x_t, y_t), t \in T$ with $x_t \in \mathbb{R}^{10X24}$ and $y_t \in \mathbb{R}$. Given the substantial size and density of the Colorado dataset, we opted to employ it for the pre-training phase of both the MAML and TL models. Subsequently, these pre-trained models underwent a fine-tuning process utilizing subsets of the ACN and both the Norway datasets. The fine-tuning was conducted using 10-day and 20-day subsets, requiring the models to make hourly load forecasts for the subsequent 10 and 20 days, respectively. This approach compelled the models to adapt to very limited historical data when making predictions about future trends.

2.3 Models

2.3.1 Neural networks

In the context of forecasting, which falls under supervised learning, a neural network (NN) can be represented as $z = f_\theta(x)$, where x represents the input, z denotes the output, and $\theta \in \mathbb{R}^n$ encompasses parameters distributed across multiple layers in the network. Typically, θ comprises weights W that signify the strength between neuronal connections in adjacent layers, as well as biases b within each layer.

These parameters are subject to optimization for a specific task $\mathcal{T} = (D, L)$, minimizing a loss function L over a designated subset of data $D = (x_t, y_t) \mid t \in T$, employed for training $D_{train} = (x_t, y_t) \mid t \in T_{train}$. x_t represents the input data and y_t stands for the corresponding labels. T encompasses the total number of samples within the dataset, and T_{train} denotes the portion of samples reserved for training. The loss function L leverages a distance metric $d()$ to capture the dissimilarity between the model outputs z and the ground truth labels y.

$$L(\theta) = \sum_{t \in T} d(z_t, y_t) = \sum_{t \in T} d(f_\theta(x_t), y_t) \tag{1}$$

For regression problems, L1 loss and mean squared error (MSE) loss are the prevalent loss functions as shown in Eqs. (2) and (3). The MSE loss is used when we want to penalize larger errors more heavily, while the L1 loss is used in the presence of outliers in the data and when the distribution of errors is not necessarily normal.

$$L_{L1}(\theta) = \frac{1}{T} \sum_{t \in T} |f_\theta(x_t) - y_t| \tag{2}$$

$$L_{MSE}(\theta) = \frac{1}{T} \sum_{t \in T} (f_\theta(x_t) - y_t)^2 \tag{3}$$

The optimal parameters θ^* are achieved through the iterative application of the back-propagation algorithm, as depicted in Eq. (4), that minimizes the loss function across the training data. Here, α represents the learning rate, and $\nabla_\theta L()$ denotes the gradient of the loss function.

$$\theta \leftarrow \theta - \alpha \nabla_\theta L(\theta) \tag{4}$$

2.3.2 LSTM

Long Short-Term Memory (LSTM) networks, an extension of NNs are designed to handle long-term dependencies [4]. The fundamental components of LSTM units encompass a cell, an input gate, an output gate, and a forget gate, each playing a distinct role in information processing [16]. In the description below, the symbols W denote weights, while b represents biases. x_t represents the current input and h_{t-1} is the historical output. A visual representation of the various LSTM network components is provided in Fig. 5.

The functioning of an LSTM network can be explained through its individual components:

Forget Gate (f_t): The forget gate determines the information discarded. It operates on x_t and h_{t-1} using weights $W_{f,x}$ and $W_{f,h}$, followed by the application of the sigmoid function $\sigma()$ which yields an output in the range (0,1).

$$f_t = \sigma(W_{f,x}x_t + W_{f,h}h_{t-1} + b_f) \tag{5}$$

Input Gate (i_t): The input gate determines what information should be retained in memory. Weights $W_{i,x}$ and $W_{i,h}$ associated with the x_t and h_{t-1}, facilitate this decision.

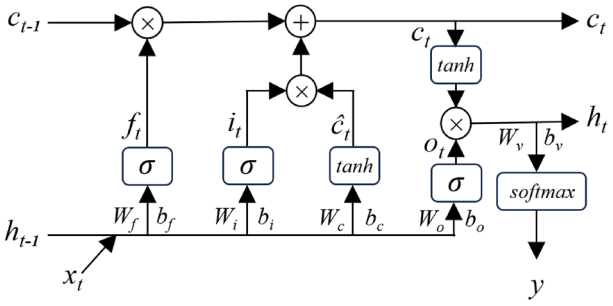


Fig. 5 Structure of an LSTM network highlighting the various gates, weights, biases and operations

$$i_t = \sigma(W_{i,x}x_t + W_{i,h}h_{t-1} + b_i) \tag{6}$$

Current Cell State Estimate ($\hat{C}_t$): A hyperbolic tangent function *tanh*(), which generates outputs within the range (-1,1), is employed to compute the estimate of the current cell state. This estimation relies on parameters $W_{c,x}$, $W_{c,h}$, and b_c.

$$\hat{C}_t = tanh(W_{c,x}x_t + W_{c,h}h_{t-1} + b_c) \tag{7}$$

Current Cell State (C_t): C_t is updated using the previous cell state C_{t-1}, $\hat{C}_t$, and the outputs of f_t and i_t.

$$C_t = f_t \times C_{t-1} + i_t \times \hat{C}_t \tag{8}$$

Output Gate (o_t): The output gate interacts with x_t and h_{t-1} through weights $W_{o,x}$ and $W_{o,h}$.

$$o_t = \sigma(W_{o,x}x_t + W_{o,h}h_{t-1} + b_o) \tag{9}$$

Current Output (h_t): h_t is generated by combining the results of the o_t and C_t.

$$h_t = o_t \times tanh(C_t) \tag{10}$$

Final Output (y): The final output y is obtained by applying a softmax function to the product of weights W_y and h_t.

$$y = softmax(W_y h_t + b_y) \tag{11}$$

LSTM models require access to substantial amounts of historical data to generate meaningful results. This can pose a challenge when dealing with newer or smaller hubs with limited data availability.

2.3.3 Transfer learning

Transfer Learning (TL) is designed to address challenges arising from limited training data. The TL process begins by initially pre-training models on a larger, separate dataset. These models are then fine-tuned through training on the target dataset, allowing them to extract general patterns from the larger dataset and adapt this knowledge to the smaller, intended dataset, as outlined in [6]. TL is employed to enhance learning in a target task, denoted as $\mathcal{T}_T = (D_T, L_T)$, by leveraging knowledge from a different yet related task, $\mathcal{T}_S = (D_S, L_S)$.

The optimal parameters θ_S, obtained from a NN trained on $D_{S,train}$ and validated on $D_{S,test}$, serve as the initial weights and biases for the NN model intended for the target task $\mathcal{T}_T$ as shown in Algorithm 1. Subsequently, the TL model undergoes fine-tuning by training on $D_{T,train}$ and is evaluated on $D_{T,test}$ to obtain optimal parameters θ_T specific to the target task $\mathcal{T}_T$.

It is essential to acknowledge a primary limitation of TL models. Fine-tuning may not fully mitigate the disparities in datasets and trends, potentially resulting in sub-optimal performance when transitioning from the source to the target task.

Algorithm 1 Transfer Learning (TL)

Require: $\mathcal{T}_S = (D_S, L_S)$: Source task
Require: α: Step size hyperparameter
1: Randomly initialize θ
2: **while** not done **do**
3: Batch load data (x_i, y_i) from $D_{S,train}$
4: Evaluate $\nabla_\theta L_{\mathcal{T}i}(f_\theta)$ using 2 or 3
5: Update $\theta \leftarrow \theta - \alpha \nabla_\theta L_{\mathcal{T}i}(f_\theta)$
6: **end while**
7: Use optimal θ_S as initial parameters for fine-tuning

2.3.4 Model-agnostic meta-learning

Model-Agnostic Meta-Learning (MAML) represents a recent advancement within the realm of Deep Learning (DL) that offers a solution to the limitations of conventional DL and TL, as introduced in [9]. In the context of meta-learning, specifically MAML, a large dataset is initially divided into numerous meta-tasks, each comprising only a limited number of data points. The model is subsequently exposed to these tasks, essentially becoming proficient in "learning to learn". This means that instead of directly updating its parameters based on an extensive external dataset, the model focuses on acquiring the ability to efficiently adjust its parameters in order to maximize its performance when faced with new tasks.

In the case of time series data, following the generation of features and time windows as outlined in Sect. 2.2, the dataset $D = (x_n, y_n)$ is partitioned into N tasks denoted as $\mathcal{T}_i, i \in (1, N)$ that each contain the same number of time windows $2l$. $2l$ consecutive time windows with a random starting point are selected for each task. Alternate time windows in each task $\mathcal{T}_i$ are segregated into exactly l support sets trained in an inner loop and l query sets trained in the outer loop. This is because alternate time windows are temporally correlated [17].

The NN model utilized for MAML is initialized with random parameters θ. The time windows $(x_{i,k}, y_{i,k}), k \in (1, l)$ in the support set of each task $\mathcal{T}_i^{sup}$ are used to iteratively update temporary parameters θ_i' by using θ and the gradient of the loss function as expressed in Eq. (12) for a number of adaptation steps set by the user. α is the inner loop learning rate.

$$\theta_i' = \theta - \alpha \nabla_\theta L_{\mathcal{T}_i^{sup}}(f_\theta) \tag{12}$$

Following this parameter update, the meta-learning rate β, parameters θ_i' and their gradients as well as time windows $(x_{i,j}, y_{i,j}), j \in (1, l)$ from the query set of each task $\mathcal{T}_i^{que}$ are employed to adjust θ, as demonstrated in Eq. (13).

$$\theta \leftarrow \theta - \beta \sum_{i=1}^{N} \nabla_\theta L_{\mathcal{T}_i^{que}}\left(f_{\theta_i'}\right) \tag{13}$$

The model, now meta-trained, can be fine-tuned on smaller datasets, allowing it to adapt quickly and learn effectively without being hampered by data scarcity or external dataset influences. The entire process is shown in Algorithm 2.

Algorithm 2 Model-Agnostic Meta-Learning (MAML)

Require: Tasks $\mathcal{T}_i, i \in (1, N)$: Derived from dataset D
Require: α, β: Step size hyperparameters
Randomly initialize θ
while not done **do**
  Sample batch of tasks $\mathcal{T}_i$
  for all $\mathcal{T}_i$ **do**
    Sample data $(x_{i,k}, y_{i,k}), k \in (1, l)$ from support set $\mathcal{T}_i^{sup}$
    Evaluate $\nabla_\theta L_{\mathcal{T}i}(f_\theta)$ using equation 2 or 3
    Compute temporary parameters with gradient descent using equation 12
    Sample data $(x_{i,j}, y_{i,j}), j \in (1, l)$ from query set $\mathcal{T}_i^{que}$ for meta-update
  end for
  Update θ using equation 13
end while
Use θ as initial parameters for fine-tuning

2.3.5 Classic machine learning

In order to justify the use of complex deep learning experiments, it is essential to compare performances with classical machine learning (CML) models. We utilize two algorithms, namely Random Forests (RF) and K-Nearest-Neighbors (KNN) due to their popularity, simplicity, and known efficacy at regression problems [18]. Other popular models such as Support Vector Machines were also included in experimentation, but their performances were consistently poor and are thus not included in the results.

The basic unit of a RF is a decision tree. A Decision Tree (DT) is a machine learning model that splits features into branches of a tree. The tree starts with a root node, representing a decision or a test on a specific feature. From there, it branches into child nodes, each corresponding to different outcomes of that decision. This branching continues until the final decisions are reached, expressed as leaf nodes. The splitting criterion, in our case, was MSE as described in Eq. 3.

RF models are a collection of DTs, each trained on a different randomly divided subset of the data. Each DT then predicts the outcome, and the final prediction is based on the average prediction of the ensemble of DTs. This ensemble approach helps to reduce overfitting.

In KNN regression, each data point is positioned in a multidimensional space according to its features. KNN locates the k-nearest instances in this space to the input data point in question via euclidean distance. The predicted value is calculated as the average of the target values of these neighbors as shown in Eq. 14. We selected *k* as 5.

$$\hat{y}(x) = \frac{1}{k} \sum_{i=1}^{k} y_i \tag{14}$$

Here, $\hat{y}(x)$ is the predicted value for the input x, k is the number of nearest neighbors considered and y_i are the target values of the k nearest neighbors to the input x.

2.3.6 Experimental setup

In our experimental setup, we employ LSTM networks as the foundational architecture. We configure an LSTM network comprising of two LSTM layers, each housing 48 neurons. This is followed by a Dense layer with 48 neurons and culminating in a single-neuron output layer. To optimize the network, we employ an RMSprop optimizer with a learning rate (α) of 0.001. Our experiments are conducted using Pytorch and the learn2learn library [19] in Python, hosted on Google Colab, which provides ample resources including 12.7 GB of RAM and an NVIDIA Tesla T4 GPU. Multiple variations of the above network were trained and tested, including number of layers, number of neurons in each layer, number of LSTM layers, bidirectionality, and learning rate. The settings mentioned were selected because they provided the best results.

For our meta-learning approach, we commence by partitioning the Colorado dataset into 500 tasks (N), with each task encompassing 200 consecutive data points (2 l). The number of tasks and data points per task were selected on the basis of performance by testing each of them in steps of 50 starting from 50 until 500. The LSTM network is encapsulated within a MAML module with a learning rate (β) of 0.05. We create two distinct models, one trained with L1 loss and the other with MSE loss. Our training process consists of 50 iterations. Within each iteration, the model encounters 32 tasks, where each task is further divided into query and support sets. Specifically, every alternate 24-hour input time window serves as a support set, while its subsequent time window serves as the query set. The model adapts to the support sets in the inner loop and evaluates its performance on the query sets, iteratively updating its parameters to enhance its performance across tasks. In each iteration, a new copy of the MAML model is employed. This practice involves cloning the original model before adaptation, effectively isolating changes made during adaptation to a specific task ensuring proper generalization.

Upon completing the meta-learning phase, we create fresh LSTM models, initializing them with the same parameters θ as the meta-trained LSTM. These models are then subjected to fine-tuning on the remaining datasets (ACN, Norway-shared, and Norway-private). We select 20-day and 40-day subsets from these datasets, splitting them evenly into training and testing sets. Fine-tuning is carried out over 50 iterations, employing L1 loss as the objective function.

In contrast, our TL approach involves training LSTM models with the previously described architecture directly on the entire Colorado dataset. Two separate models are trained-one with L1 loss and the other with MSE loss-for 50 iterations. Subsequently, these models are fine-tuned on the ACN and Norway datasets in a manner consistent with the process outlined above.

For traditional DL, where no pretraining is involved, models are directly trained on the first halves of the selected subsets for 50 iterations using L1 loss and then evaluated on the remaining halves. The CML models are also directly trained on the same portions of the data after flattening the input vector. The evaluation metrics include mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R2) scores. To ensure robustness, each experiment is conducted five times, and the average values of all metrics are reported.

3 Results

This section is divided into 3 subsections pertaining to each of the datasets under investigation. Tables 1, 2 and 3 provide the metrics for the seven EV charging load forecasting models : TL with MSE and L1 loss, MAML with MSE and L1 loss, DL with L1 loss, RF and KNN.

3.1 ACN dataset

Analyzing the ACN dataset results presented in Table 1, it becomes evident that TL L1 exhibits the most favorable performance, particularly in terms of MAE and RMSE. Notably, TL L1's RMSE score is surpassed only by TL MSE in two subsets of the ACN dataset considered. In contrast, DL consistently lags behind the performance of competing models across all metrics. With the exception of the first and third subsets wherein RF performs competitively, CML models are well-beaten by the deep learning models.

Furthermore, when comparing L1 variants to their MSE counterparts within the MAML and TL models, it becomes apparent that L1 variants outshine their MSE counterparts. Additionally, examining the R-squared (R2) scores reveals that TL models excel in capturing and adhering to the underlying trends in the data.

Analyzing the trends depicted in Figs. 6 and 7, we can evaluate how well each model aligns with the actual hourly load patterns. It is evident that, in general, models tend to exhibit suboptimal performance during peak load periods, occasionally either overestimating or underestimating the peak values. Interestingly, models display a higher level of accuracy in predicting periods of zero load, with some exceptions noted in the case of DL and MAML L1, as illustrated in Fig. 6. From both figures, we see that the CML models perform very poorly during zero-load hours. While they are able to understand the general trend, they predict the trend to continue even when there is no actual load.

3.1.1 Norway shared charger dataset

Examining the Norway shared charger data presented in Table 2, it is clear that MAML demonstrates the most robust performance, specifically excelling in terms of MAE and RMSE. Interestingly, in this dataset, DL manages to outperform TL models, yet it still falls short when compared to the performance achieved by MAML

Table 1 Mean MAE, RMSE and R2 scores of the forecasting models on the ACN dataset

Specification		Mean scores		
Subset	Model	MAE	RMSE	R2
18th September 2019, 10 Days	TL MSE	1.48	***2.97***	***0.83***
	TL L1	***1.40***	3.03	0.82
	MAML MSE	1.58	3.17	0.81
	MAML L1	1.53	3.19	0.81
	DL	1.95	3.53	0.75
	RF	1.51	2.99	0.74
	KNN	2.27	3.93	0.54
9th November 2020, 10 Days	TL MSE	0.18	0.76	0.72
	TL L1	***0.14***	***0.70***	***0.76***
	MAML MSE	0.20	0.78	0.70
	MAML L1	0.18	0.79	0.69
	DL	0.19	0.78	0.70
	RF	0.34	0.93	0.16
	KNN	0.5	1.27	0.01
18th September 2019, 20 Days	TL MSE	1.52	3.42	0.68
	TL L1	***1.37***	3.37	0.69
	MAML MSE	1.47	3.38	0.69
	MAML L1	1.47	3.38	0.69
	DL	1.46	3.39	0.69
	RF	1.51	***3.21***	0.68
	KNN	1.93	3.79	0.54
9th November 2020, 20 Days	TL MSE	0.30	***0.96***	***0.72***
	TL L1	***0.29***	1.00	0.70
	MAML MSE	0.34	1.05	0.67
	MAML L1	0.31	1.03	0.68
	DL	0.33	1.03	0.67
	RF	0.39	1.10	0.52
	KNN	0.8	1.60	0.09

The values in bold and italics represent the best metrics achieved by the models for the subset of data in consideration

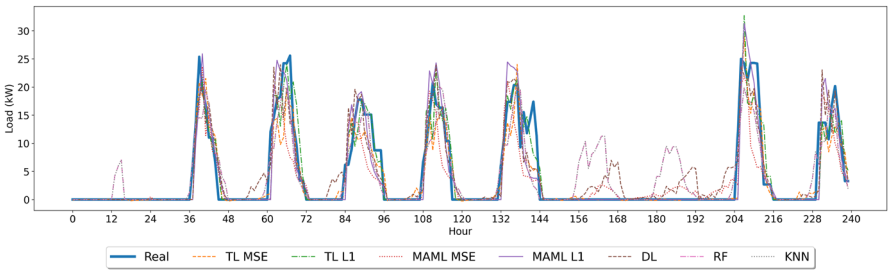


Fig. 6 Results from the forecasting models compared to the true load value on a subset of the ACN dataset starting from September 18, 2019 for 10 days

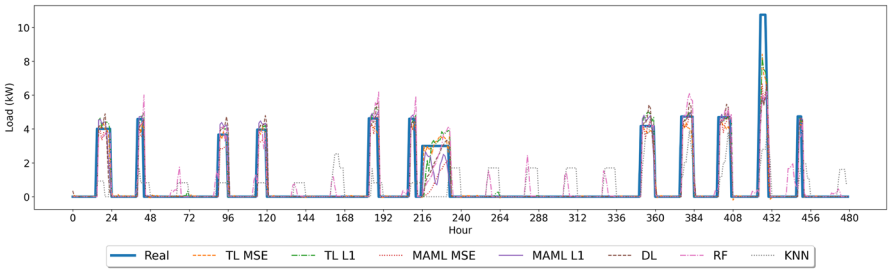


Fig. 7 Results from the forecasting models compared to the true load value on a subset of the ACN dataset starting from November 9, 2020 for 20 days

Table 2 Mean MAE, RMSE and R2 scores of the forecasting models on the Norway shared charger dataset

Specification		Mean scores		
Subset	Model	MAE	RMSE	R2
	TL MSE	0.96	2.38	0.63
	TL L1	0.91	2.39	0.63
	MAML MSE	0.91	2.38	0.63
21st February 2019, 10 Days	MAML L1	***0.89***	***2.32***	***0.65***
	DL	0.92	2.40	0.62
	RF	1.18	2.48	0.32
	KNN	2.28	4.3	−0.32
	TL MSE	3.77	5.64	0.59
	TL L1	3.66	5.50	0.61
	MAML MSE	***3.30***	***5.06***	***0.67***
17th September 2019, 10 Days	MAML L1	3.47	5.29	0.64
	DL	3.60	5.55	0.61
	RF	3.51	4.94	0.46
	KNN	5.88	8.36	−1.84
	TL MSE	1.01	2.43	0.59
	TL L1	1.03	2.44	0.59
	MAML MSE	0.93	2.35	0.62
21st February 2019, 20 Days	MAML L1	***0.88***	***2.25***	***0.65***
	DL	0.97	2.41	0.60
	RF	1.24	2.31	0.49
	KNN	2.17	3.83	−0.23
	TL MSE	3.70	5.64	0.65
	TL L1	3.38	5.50	0.67
	MAML MSE	***3.23***	***5.38***	***0.68***
17th September 2019, 20 Days	MAML L1	3.38	5.52	0.67
	DL	3.33	5.45	0.67
	RF	4.29	5.68	0.33
	KNN	5.93	8.18	−1.03

The values in bold and italics represent the best metrics achieved by the models for the subset of data in consideration

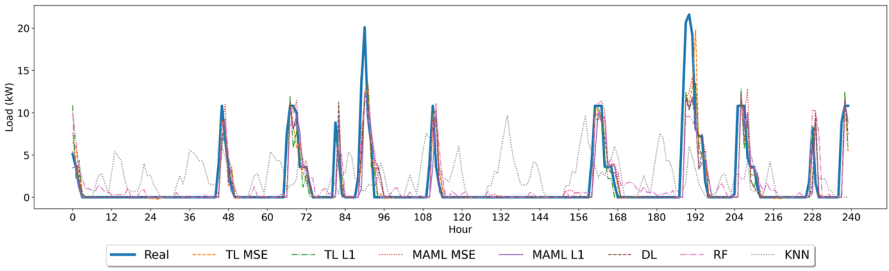


Fig. 8 Results from the forecasting models compared to the true load value on a subset of the Norway shared charger dataset starting from February 21, 2019 for 10 days

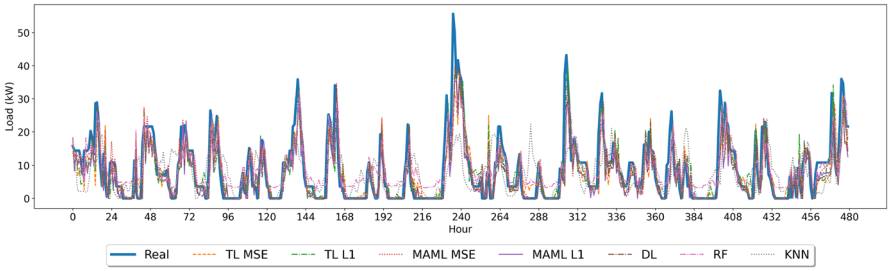


Fig. 9 Results from the forecasting models compared to the true load value on a subset of the Norway shared charger dataset starting from September 17, 2019 for 20 days

models. CML models expectedly underperform, with KNN showing negative R2 scores. In the second subset of the dataset, the MAE exhibited by DL is 10% higher than that of MAML MSE but 5% lower than that of TL MSE. Additionally, the R2 scores affirm that MAML models tend to make predictions that closely align with the actual data trends, demonstrating their superior predictive capability.

From Figs. 8 and 9, we can assess how each model fits the real hourly load trend. We can see that the models under-perform at the two peaks greater than 15 kW in Fig. 8, and the CML models continue to predict loads even when there are none, as discussed previously. Additionally, Fig. 8 displays an interesting phenomenon—models appear to arrive with a delay of one timestep with respect to the peaks, suggesting that they are under-confident in predicting the starting edges of the peak periods.

3.1.2 Norway private charger dataset

From the data outlined in Table 3, it becomes evident that TL L1 demonstrates the most robust performance across all subsets examined within the Norway private charger dataset. Conversely, DL exhibits relatively poorer performance in most instances.

Table 3 Mean MAE, RMSE, and R2 scores of the forecasting models on the Norway private charger dataset

Specification		Mean scores		
Subset	Model	MAE	RMSE	R2
	TL MSE	2.8	4.95	0.51
	TL L1	***2.61***	***4.93***	***0.58***
	MAML MSE	3.1	6.35	0.2
21st February 2019, 10 Days	MAML L1	2.98	6.18	0.24
	DL	3.12	6.72	0.18
	RF	2.80	5.60	0.14
	KNN	3.66	7.28	− 0.10
	TL MSE	5.01	7.73	0.59
	TL L1	***4.75***	***6.96***	***0.69***
	MAML MSE	6.18	9.23	0.46
17th September 2019, 10 Days	MAML L1	5.74	8.18	0.58
	DL	6.24	9.34	0.45
	RF	4.85	6.65	0.61
	KNN	6.24	9.34	0.45
	TL MSE	3.10	5.65	0.58
	TL L1	***2.98***	***5.64***	***0.58***
	MAML MSE	3.92	7.01	0.35
21st February 2019, 20 Days	MAML L1	3.39	6.38	0.45
	DL	3.59	6.66	0.41
	RF	3.23	6.28	0.43
	KNN	5.17	9.01	0.12
	TL MSE	5.56	8.95	0.69
	TL L1	***5.39***	***8.28***	***0.74***
	MAML MSE	5.61	8.45	0.72
17th September 2019, 20 Days	MAML L1	5.39	8.35	0.73
	DL	5.70	8.93	0.69
	RF	5.50	8.57	0.46
	KNN	7.36	10.21	−0.31

The values in bold and italics represent the best metrics achieved by the models for the subset of data in consideration

For example, consider the MAE scores: TL L1 achieves scores that are 20% and 31% lower than those attained by DL for the first and second subsets within the dataset. Moreover, TL L1 outperforms MAML L1 (the better-performing MAML) by approximately 15% and 21% for the same subsets. CML models expectedly underperform again. Additionally, it's worth noting that both MAE and RMSE values for all models within the Norway private charger dataset are higher when compared to those of the Norway shared charger dataset and the ACN dataset. This discrepancy might be attributed to the Colorado dataset containing charging transactions acquired from shared chargers that might exhibit similar patterns to Norway shared

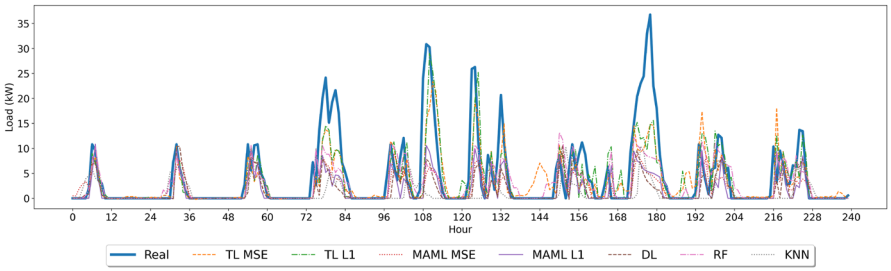


Fig. 10 Results from the forecasting models compared to the true load value on a subset of the Norway private charger dataset starting from February 21, 2019 for 10 days

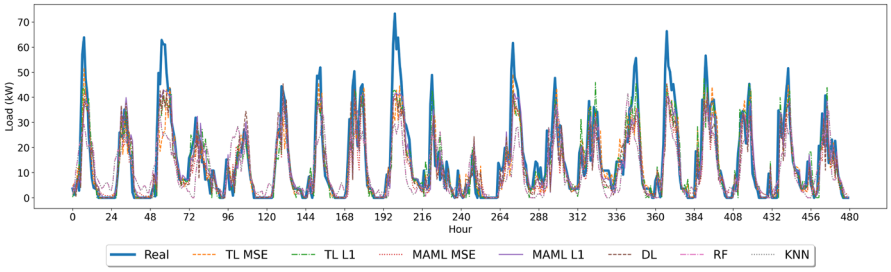


Fig. 11 Results from the forecasting models compared to the true load value on a subset of the Norway private charger dataset starting from September 17, 2019 for 20 days

charger dataset and the ACN dataset (shared chargers on campus). Achieving an optimal balance between private and shared charging data during pre-training may be necessary to enhance the final results.

Analyzing Figs. 10 and 11, we can evaluate how well each model aligns with the actual hourly load trends. Notably, all models, with the exception of TL L1 and TL MSE, exhibit significant underperformance. TL models struggle to accurately predict peaks exceeding 15 kW as seen in Fig. 10. However, Fig. 11 shows a relatively improved performance, though the models still face challenges in accurately predicting peaks exceeding 50 kW.

4 Discussion

Our research demonstrates the effectiveness of MAML and TL models in short-term load forecasting, particularly in scenarios with limited data availability. These models exhibit the capability to forecast accurately across diverse geographical regions, scales, and data distributions by employing a pre-training strategy on a more extensive dataset followed by fine-tuning on the smaller target dataset. Notably, both MAML and TL models exhibit competitive performance, with MAML excelling for the Norway shared charger dataset and TL performing exceptionally well for the ACN dataset as well as the Norway private charger dataset. MAML seems to perform

best when the target dataset closely matches the setting of the source dataset but this does not seem to matter too much for TL. In this work, the Colorado dataset was the only dataset used for pre-training and the data is obtained from shared chargers dispersed in Colorado which is similar to the Norway shared charger dataset with data from shared chargers in Trondheim. Achieving an optimal balance between private and shared charging data during pre-training may be necessary to enhance forecasting metrics. Across all three evaluation metrics, both MAML and TL models consistently outperform traditional Deep Learning (DL) models. Classical Machine Learning (CML) models are clearly not adequate for the complexity of the task. It's worth noting that while the differences in scores are generally significant, there may be instances where DL models outperform specific MAML and TL configurations.

We also see that models tend to exhibit suboptimal performance during peak load periods, occasionally either overestimating or underestimating the peak values. However, most models display a higher level of accuracy in predicting periods of zero load except CML models. While CML models are able to understand the general trend, they predict the trend to continue even when there is no actual load. It is also worth noting that sometimes the models appear to perform well after a delay of one timestep when the load jumps from zero to a higher value. This suggests that the models are under-confident in predicting the starting edges of the peak periods.

A few variations of the models were tested before arriving at the final models. We tried pre-training the models for MAML and TL on the same timespan as the test data i.e. if the test data was October to November 2019 then we pre-trained the models of data only from August to September 2019. The results were poor compared to the results discussed above. Additionally, multiple variations of the models were tested by tuning various hyperparameters before settling on the model architectures utilized for forecasting.

4.1 Future research

The addition of more diverse data in the source dataset for pre-training can potentially improve forecasting accuracy. Another avenue worth exploring is the possibility of predicting individual user-specific charging loads rather than aggregate loads with aggregate load pre-training, potentially protecting privacy and enhancing forecasting precision. Additionally, the adoption of ensemble models (combination of TL and MAML) could be investigated as a means to further improve forecasting metrics.

In the realm of medium and heavy-duty electric vehicles EVs, MAML and TL find valuable applications beyond forecasting aggregate loads. They prove beneficial for predictive maintenance, onboard diagnostics, and adapting to various routes, drivers, and weather conditions. Predictive maintenance models have the capacity to adjust to the unique usage patterns and wear and tear experienced by each individual truck. MAML aids in the generation of optimal driving decisions, facilitating rapid adaptation to diverse driving conditions, routes, and cargo loads.

Mandates in specific regions, such as California, may offer access to extensive early-stage data, which can be harnessed for pre-training. Subsequently, these

models can undergo fine-tuning for future adopters. This process contributes to the planning of advanced high-power charging infrastructure and enhances driving and energy efficiency.

Acknowledgements We are grateful to the sponsorship received from the following Grants 69763, 77739, and 45779 to the department of Mechanical and Aerospace Engineering, UCLA.

Author Contributions SN: conceptualization, investigation, formal analysis, methodology, data curation, software, validation, visualization, writing—original draft preparation, KN: conceptualization, investigation, methodology, data curation, software, validation, visualization, writing—review and editing, CZ: conceptualization, investigation, project administration, writing—review and editing, RG (Rohan Gowda): software, validation, visualization, writing—review and editing, RG (Rajit Gadh): supervision, project administration, funding acquisition, writing—review and editing.

Availability of data and materials The publicly available data sources have been cited.

Declarations

Conflict of interest The authors declare that there is no potential conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Balasubramanian, S., Balachandra, P.: Effectiveness of demand response in achieving supply-demand matching in a renewables dominated electricity system: A modelling approach. Renew. Sustain. Energy Rev. **147**, 111245 (2021)
2. Huang, X., Wu, D., Boulet, B.: Ensemble learning for charging load forecasting of electric vehicle charging stations. In: 2020 IEEE Electric Power and Energy Conference (EPEC), pp. 1–5 (2020). IEEE
3. Sadeghianpourhamami, N., Refa, N., Strobbe, M., Develder, C.: Quantitive analysis of electric vehicle flexibility: a data-driven approach. Int. J. Electr. Power Energy Syst. **95**, 451–462 (2018)
4. Hewamalage, H., Bergmeir, C., Bandara, K.: Recurrent neural networks for time series forecasting: current status and future directions. Int. J. Forecast. **37**(1), 388–427 (2021)
5. Lee, E., Rhee, W.: Individualized short-term electric load forecasting with deep neural network based transfer learning and meta learning. IEEE Access **9**, 15413–15425 (2021)
6. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. J. Big Data **3**(1), 1–40 (2016)
7. Forootani, A., Rastegar, M., Zareipour, H.: Transfer learning-based framework enhanced by deep generative model for cold-start forecasting of residential ev charging behavior. IEEE Trans. Intell. Veh. (2023)
8. Banda, P., Bhuiyan, M.A., Hasan, K.N., Zhang, K.: Assessment of hybrid transfer learning method for forecasting ev profile and system voltage using limited ev charging data. Sustain. Energy Grids Netw. **36**, 101191 (2023)
9. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning, pp. 1126–1135 (2017). PMLR

10. He, Y., Luo, F., Ranzi, G.: Transferrable model-agnostic meta-learning for short-term household load forecasting with limited training data. IEEE Trans. Power Syst. **37**(4), 3177–3180 (2022)
11. Xiao, X., Mo, H., Zhang, Y., Shan, G.: Meta-ann—a dynamic artificial neural network refined by meta-learning for short-term load forecasting. Energy **246**, 123418 (2022)
12. Lee, Z.J., Li, T., Low, S.H.: Acn-data: analysis and applications of an open ev charging dataset. In: Proceedings of the Tenth ACM International Conference on Future Energy Systems, pp. 139–149 (2019)
13. Sørensen, Å.L., Lindberg, K.B., Sartori, I., Andresen, I.: Analysis of residential ev energy flexibility potential based on real-world charging reports and smart meter data. Energy Build. **241**, 110923 (2021)
14. Boulder Open Data, C.: Electric vehicle charging station data. https://open-data.bouldercolorado.gov/datasets/95992b3938be4622b07f0b05eba95d4c_0/explore
15. Sørensen, Å.L., Lindberg, K.B., Sartori, I., Andresen, I.: Residential electric vehicle charging datasets from apartment buildings. Data Brief **36**, 107105 (2021)
16. Dong, Y., Chen, Y., Zhao, X., Huang, X.: Short-term load forecasting with distributed long short-term memory. In: 2023 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp. 1–5 (2023). IEEE
17. Arango, S.P., Heinrich, F., Madhusudhanan, K., Schmidt-Thieme, L.: Multimodal meta-learning for time series regression. In: Advanced Analytics and Learning on Temporal Data: 6th ECML PKDD Workshop, AALTD 2021, Bilbao, Spain, September 13, 2021, Revised Selected Papers 6, pp. 123–138 (2021). Springer
18. Fernández-Delgado, M., Sirsat, M.S., Cernadas, E., Alawadi, S., Barro, S., Febrero-Bande, M.: An extensive experimental survey of regression methods. Neural Netw. **111**, 11–34 (2019)
19. Arnold, S.M., Mahajan, P., Datta, D., Bunner, I., Zarkias, K.S.: learn2learn: a library for meta-learning research (2020). arXiv preprint arXiv:2008.12284

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Shashank Narayana Gowda[1] · Keshav Nath[2] · Chen Zhang[1] · Rohan Gowda[3] · Rajit Gadh[1]

✉ Shashank Narayana Gowda
shashank07@ucla.edu

Keshav Nath
nathkeshav14@gmail.com

Chen Zhang
chenzhang96@g.ucla.edu

Rohan Gowda
rsg3418@g.coppellisd.com

Rajit Gadh
gadh@ucla.edu

[1] Smart Grid Energy Research Center, University of California Los Angeles, Los Angeles, CA 90095, USA

[2] Department of Mechanical Engineering, Delhi Technological University, New Delhi, Delhi 110042, India

[3] Coppell High School, Coppell, TX 75019, USA

Plasmonics
https://doi.org/10.1007/s11468-024-02396-0

CORRESPONDENCE

Synthesis of Graphene Quantum Dots (GQDs) from Paddy Straw for Bilirubin Detection

Isha[1] · Aneesha[1] · Mohan Singh Mehata[1]

Received: 8 May 2024 / Accepted: 6 June 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

This work reports the synthesis of blue-emitting graphene quantum dots (GQDs) with an average diameter of 3.8 ± 0.5 nm from paddy straw, a sustainable biomass resource, via hydrothermal synthesis. These GQDs demonstrate excellent performance in bilirubin (BR) detection. The GQD photoluminescence (PL) intensity exhibits a proportional decrease with increasing BR concentration, indicating efficient quenching. The limit of detection for BR reaches a low value of 87.9 nM, highlighting the high sensitivity and selectivity of the GQD-based sensor. The observed quenching likely arises from a combined mechanism involving static quenching due to GQD-BR complex formation, inner filter effect (IFE), and Förster resonance energy transfer facilitated by spectral overlap.

Keywords Biomass waste · Paddy straw · Graphene quantum dots · Bilirubin · Forster resonance energy transfer

Introduction

Carbon nanostructures have collected all important attention repaid to their striking mechanical vigorous, electrically conductive abilities, and noteworthy chemical and thermal qualities [1]. In particular, graphene, known as the principal form of carbon graphite, contains only one layer of carbon atoms joined together by coinciding sp^2 hybrid bonds [2]. Unlikely graphene, GQDs are non-zero band gap materials [3] that exhibit their strong quantum confinement [4] and edge effects [5]; for this reason, GQDs possess remarkable optical and electrical properties. They serve important properties such as strong water solubility, low cost, stable photoluminescence (PL), tunable bandgap, good compatibility, and low toxicity [6–8]. GQDs possess their unique inner structure and surface chemical groups [9]. GQDs normally have features that drop below a diameter of 20 nm [10]. Because of its small size, GQDs carry a declaration for biomedical applications in comparison with graphene or graphene oxide (GO) [11, 12].

Biomass waste, described from its green, affordable, abundant, easily accessible, and carbon-rich nature, introduces a promising option as a precursor for creating GQDs. The management and renewal of biomass waste presents an important confrontation in today's world. The effective use of biomass waste not only helps in reusable obstacles but also gives a low-budget source for GQDs synthesis. This plan considered using biomass waste as a low-budget and sustainable precursor [13]. Rice is one of the highest-expanded crops globally [14]; however, agricultural waste for example paddy straw is a usual agricultural residue that consists of generous carbon content because of the existence of cellulose and lignin [15]. The maximum paddy straw manufactured is currently inclined through open field burning. This procedure leads to air pollution, energy wastage, and the occupation of landfill areas [16]. Constructively modifying paddy straw biomass waste into valuable products can capitulate essential economic and environmental benefits [17].

GQDs synthesis techniques are often divided into two distinct categories: top-down and bottom-up methods. Bottom-up techniques utilize detailed effective steps and particulate organic elements, imposing obstacles for optimizing circumstances [11]. On the other hand, top-down techniques include breaking down and separating easily available bulk graphene-based materials, like graphite, under the worst circumstances. These procedures typically consist of different steps involving strong oxidizing agents, concentrated

✉ Mohan Singh Mehata
msmehata@gmail.com

[1] Laser Spectroscopy Laboratory, Department of Applied Physics, Delhi Technological University, New Delhi 110042, India

Published online: 28 June 2024

acids, and elevated temperatures [10]. Top-down techniques include sol–gel method [18], hydrothermal method [19], electrochemical exfoliation [20], chemical oxidation [21], nanolithography [22], microwave-assisted [23], and ultrasonic shearing [24]. The PL can be readily monitored by simply altering the dimension and surface chemistry of GQDs. GQDs originated from paddy straw that can be dispersed in water with strong PL and excellent biocompatibility, presenting better capability for use in biological and medical fields [24]. GQDs have been employed to detect BR, which is a yellow pigment with a structural formula, as shown in Fig. 1, produced during the instinctive breakdown of red blood cells (RBCs) as part of the body's cleansing process [25]. It is an important component of bile, an essential digestive liquid the liver produces. Uplifted steps of unconjugated BR in the blood are important signs of many strict situations such as jaundice and hyperbilirubinemia [26, 27]. The ongoing analysis proves a hydrothermal method for synthesizing GQDs, engaging paddy straw as an economical raw material. The GQDs produced exhibit high and particular sensitivity to BR, making them potentially useful for BR detection purposes.

Experimental Sections

Materials

The agricultural waste of paddy straw was gathered from Pundri village in the Kaithal district of Haryana, India, and kept in dry conditions at room temperature to serve as a precursor for synthesizing GQDs. The analytes bilirubin, urea, uric acid, fructose, glucose, 7-azaindole (AI), ethylenediamine-tetra acetic acid (EDTA), and bis-(trimethylsilyl) acetamide (BSA) were all acquired from CDH chemicals.

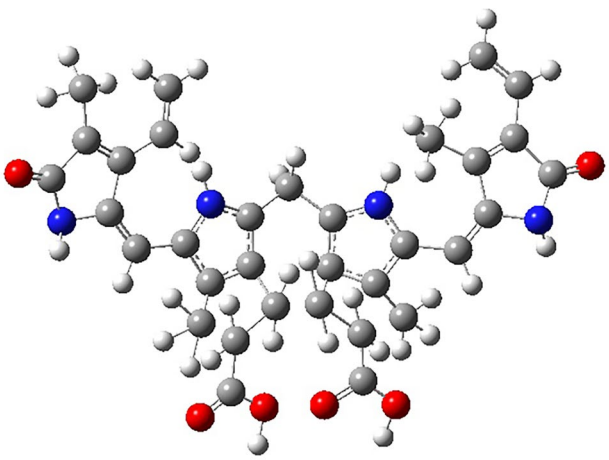


Fig. 1 Structure of bilirubin

Deionized water (DI) and hydrochloric acid (HCL) were utilized to prepare solutions for the analysis.

Preparation of Graphene Quantum Dots from Paddy Straw

A specimen of paddy straw weighing 2.0 g was taken, washed thoroughly with DI water, dried, and then ground into fine powder using mechanical crushing. Subsequently, the powder was heated at 50 °C for 3 h after being washed with 0.15 M HCl. Further, in 25 mL of water, 150 mg of crushed paddy straw was mixed and then placed in an oven for 7 h in a Teflon-lined autoclave at 160 °C. After cooling, the solution was filtered with a pore size of filter to be 11 µM, and additionally, the GQDs were further purified by centrifugation at 8000 rpm for 15 min, which separates them from any bigger precursor materials, byproducts, or contaminants that might be penetrated through the filter paper. This byproduct was gathered at the tube's bottom as precipitation in the form of silica. We collected the supernatant and used it for the experiment. Precipitates were thrown away. Figure 2 depicts a schematic illustration of the GQD preparation procedure.

Characterization

UV–vis spectroscopy experiments were used to analyze and study absorption spectra using a dual–beam UV/VIS/NIR spectrometer (Lambda 750, Perkin Elmer, USA) [28]. For the samples, a quartz cuvette with optical path of 10 mm was used. The Fluorolog-3 spectrofluorometer (Horiba Jobin Yvon) was utilized to record PL and PL-excitation measurements, which was configured with a photomultiplier tube (PMT), 450W Xenon, and flash lamps. Using Bruker D8 advance which is equipped with a Cu (K_α) radiation source with an accelerating voltage of 40 kV and 20 mA current, X-ray diffraction (XRD) patterns were obtained [1]. Fourier-transform infrared spectroscopy (FT-IR) was used to obtain the spectra under the frequency range 500–4000 cm^{-1} using Perkin Elmer, FTIR spectrometers. High-resolution images of the GQDs were taken by transmission electron microscopy (TEM) with an accelerating voltage of 200 kV using TALOS thermos-scientific instruments.

Results and Discussion

Structural, Morphological, and Optical Characteristics of GQDs

The XRD was utilized to study the structural property. Figure 3 displays the peak at 25.4°, which corresponds to the (002) diffraction plane in the XRD of GQDs [1, 29]. The

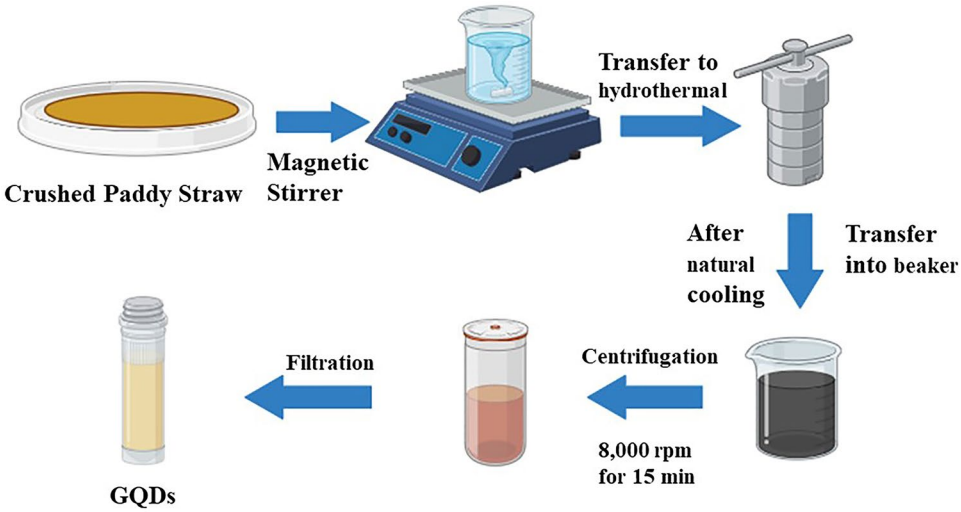


Fig. 2 Schematic representation preparation of GQDs

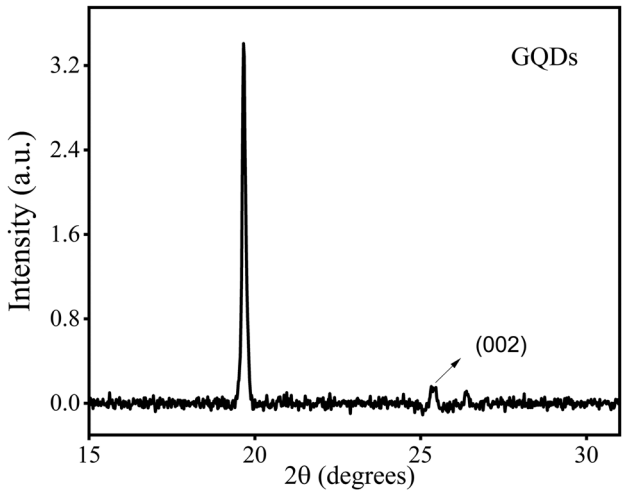


Fig. 3 XRD pattern of GQDs

highest peak in the figure may indicate the existence of residual silica ions [30, 31]. This silica may come from the sediment remaining at the bottom of the centrifuge [32].

The fabricated GQDs were also analyzed through TEM, as illustrated in Fig. 4a. To prepare the sample, drop the sample solution onto a grid of copper and then allow it to dry at room temperature. The size distribution of the GQDs, which is 3.8 ± 0.5 nm is illustrated in Fig. 4b. The observed spherical particles are fairly polydispersed.

Figure 5 depicts the FTIR spectrum of GQDs, where a wide band at 3318 cm^{-1} is assigned to the stretching vibrations of amine (N–H) and hydroxyl (–OH) groups [33]. The presence of OH and C–H linkages is revealed by the peak observed at 2112 cm^{-1}. The stretching vibrations of C–O and aromatic rings C = C, which correspond to epoxy and ether groups, respectively, are responsible for the observed peaks at 1639 and 1103 cm^{-1} [13]. Because of the abundance of oxygen-bearing functional groups, GQDs are allowed to remain stably suspended and effortlessly disperse in water.

The optical parameters of GQDs were examined using a steady-state spectroscopic technique. The UV–vis absorption

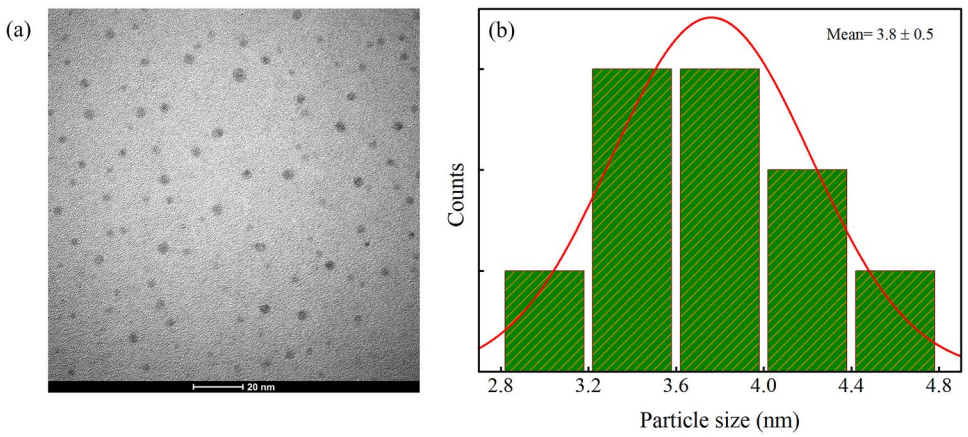


Fig. 4 TEM images (**a**) and size distribution (**b**) of GQDs

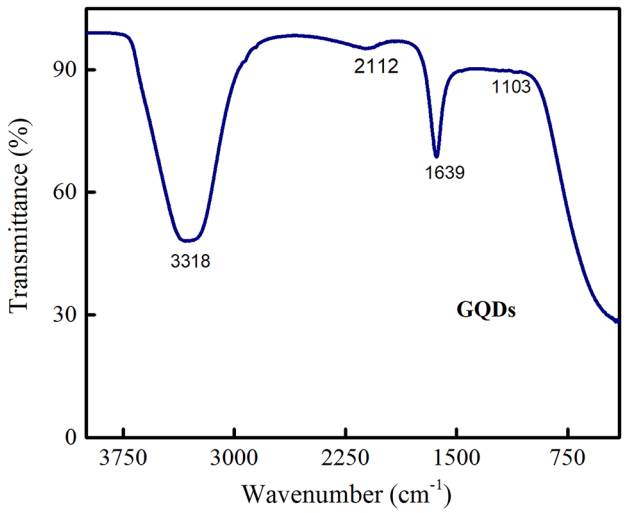


Fig. 5 FT-IR spectrum of GQDs

spectrum of GQDs is illustrated in Fig. 6a, showing a prominent absorption peak in the UV region at around 278 nm. The intense and sharp peak of absorption below 300 nm exhibits the π-π* transitions of the C = C aromatic bonds, indicating the involvement of sp^2 hybridization domains. Also, the functionalized oxygen group on the surface of GQDs generates an absorption associated with the n-π* transition of the C = O bonds, which is observable in the region of 290–320 nm [28]. From the absorption spectrum, the optical bandgap of the GQDs was determined by using a Tauc plot [34], which illustrates the relationship between $(\alpha h\nu)^{1/\gamma}$ vs. photon energy $h\nu$. The γ factor, which is reliant upon the nature of the material is ½ for direct and 2 for indirect bandgap materials, respectively [35]. Therefore, a direct bandgap of GQDs was estimated to be ~ 4.07 eV as illustrated in Fig. 6b. Upon stimulation at the excitation wavelength, the GQDs demonstrate intense PL.

The optical properties were further investigated, and photoluminescence (PL) spectra were studied across a wide range of excitation wavelengths. As illustrated in Fig. 7a, when excited at a wavelength of 370 nm, the PL spectrum exhibits a broad peak with maximum intensity at 450 nm and the position of the peak changes gradually with variations in excitation wavelengths (Fig. 7a, b). A large stokes shift is observed with excitation wavelengths [36]. The intensity of PL from the GQDs decreases as the excitation wavelength increases, resulting in a redshift. PL maximum shifted in response to variations in the excitation wavelength (Fig. 7b). The Stokes shift in wavelength offers a beneficial feature for utilizing GQDs in sensing applications [37–39]. Using quinine sulfate, which has a quantum yield (QY) of 54% as a reference, it was determined that the QY of GQDs was 3.5% using following equation.

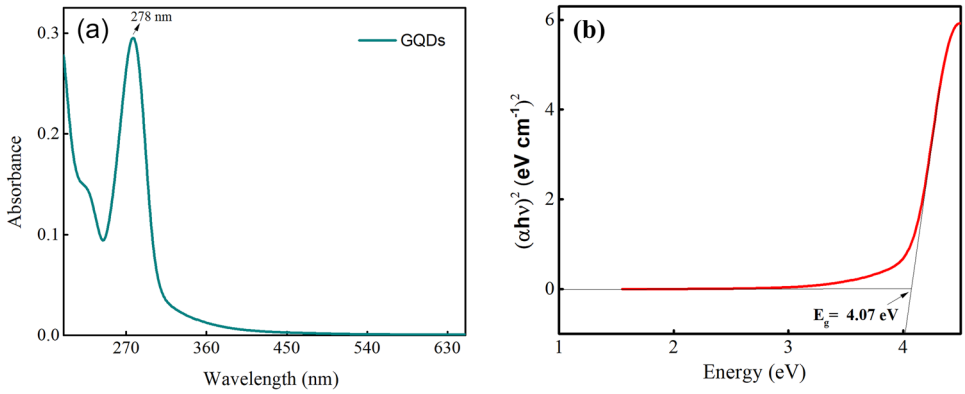


Fig. 6 UV–vis spectrum of GQDs (**a**) and Tauc plot for direct band gap GQDs (**b**)

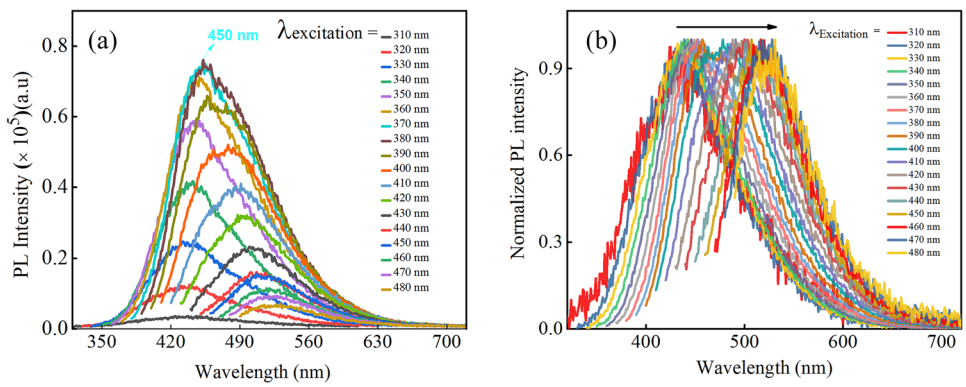


Fig. 7 Photoluminescence spectra at a range of excitation from 310 to 480 nm (**a**) and normalized PL spectra (**b**) of GQDs

$$QY_{sam} = QY_{ref}\left(\frac{PL_{sam}}{PL_{ref}}\right)\left(\frac{Abs_{ref}}{Abs_{sam}}\right)\left(\frac{\eta_{sam}}{\eta_{ref}}\right)^2$$

where QY_{ref} indicates the quantum yield of quinine sulfate (reference) and QY_{sam} indicates the quantum yield of GQDs. Abs indicates the absorbance of reference and GQDs, and the integrated areas of PL spectra are represented by PL_{sam} and PL_{ref} and η represent the refractive index of distilled water and sulfuric acid, which are used as the solvents in QDs and reference, respectively.

Sensing of Bilirubin

Preparation of Bilirubin Solution

A stock solution of bilirubin (1 mM) was prepared by dissolving the required quantity of bilirubin in water by adjusting the pH to 8 using a 1 M NaOH solution, as bilirubin does not dissolve in water, but it becomes soluble in an alkaline environment. A small portion was then taken from this stock solution, and the pH was adjusted accordingly for the experimental requirements using a 1 M HCl solution when needed. All experiments were conducted in low-light conditions.

Sensitivity and Selectivity of BR

Various concentrations of bilirubin were prepared in an ideal environment to analyze the sensing capabilities of GQDs. As illustrated in Fig. 8a, the absorption spectra of GQDs at different concentrations ranging from 0.0 to 115 µM were observed, with and without the presence of BR. Along with the absorption intensity at 278 nm, the band-edge absorption intensity at ~ 327 nm also increased. The intensity of the band-edge increased due to the n-π^* transition in GQDs. When the concentration of BR increases, the intensity of the absorption also increases and is nearly constant at the concentration of 115 µM and higher [40]. The increase in the absorption intensity indicates the interaction between GQDs and BR. This interaction might involve the formation of a stable complex due to bonding between the hydroxyl group of BR and functional groups of the GQDs in the ground state. Such complex formation between GQDs and BR potentially leads to enhanced absorption intensity at 278 nm without showing any extra peak of BR. To further strengthen the evidence for this interaction, the Benesi-Hildebrand (BH) method was employed [8]. The plot of $1/(A_0\text{-}A)$ vs. 1/conc. of BR, where A_0 represents the absorbance without BR and A represents the absorbance with BR in the range of 0–115 µM is quite linear with $R^2 = 0.99$, as depicted in the inset of Fig. 8a. The plot indicates the formation of a 1:1 complex between BR and GQDs. The calculated binding constant (K_b) is $5.15 \times 10^3\ M^{-1}$ [41]. The linear relation observed between the absorbance of GQDs and the concentration of BR signifies that GQDs can be utilized as a probe for the detection of BR.

Before bilirubin addition, the PL spectra exhibited a peak at 450 nm when excited with a wavelength of 370 nm. To analyze the PL, different concentrations of bilirubin were mixed with the GQDs solution and it has been acclaimed that if there is an increase in the concentration of bilirubin, the PL intensity of the GQDs is significantly reduced as shown in Fig. 8b. Gradually adding bilirubin aliquots led to a consistent decrease in peak intensity. At a bilirubin concentration of 115 µM, almost 90% of the intensity was quenched. While the findings suggest the formation of a ground-state complex between BR and GQDs (Fig. 9), it also leads to the quenching of GQDs PL intensity along with several possible mechanisms, such as non-fluorescent ground-state complex, inner filter effect (IFE), and Förster resonance energy transfer (FRET) [42]. The reduction in peak intensity corresponding to a specific bilirubin concentration was quantified using a Stern–Volmer (SV) plot that shows a relationship between the PL intensity ratio (I_0/I)-1 vs. bilirubin concentration, where I and I_0 represent the PL peak intensity in the presence and absence of bilirubin, respectively. Based on the plot analysis, a linear regression ($R^2 = 0.99$) was obtained within the range of 9.9–115 µM, as depicted in the inset of Fig. 8b.

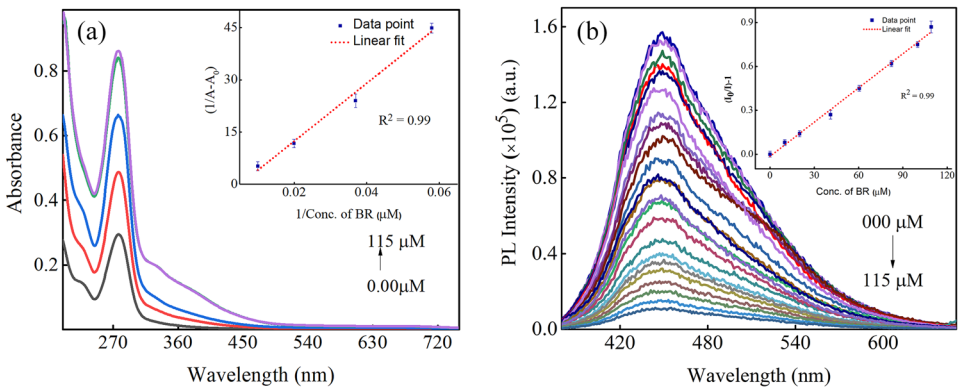


Fig. 8 Absorption spectra (**a**) and photoluminescence spectra (**b**) of GQDs in the absence and presence of multiple BR concentrations and inset of **a** represent BH plot and **b** the corresponding linear S-V plot

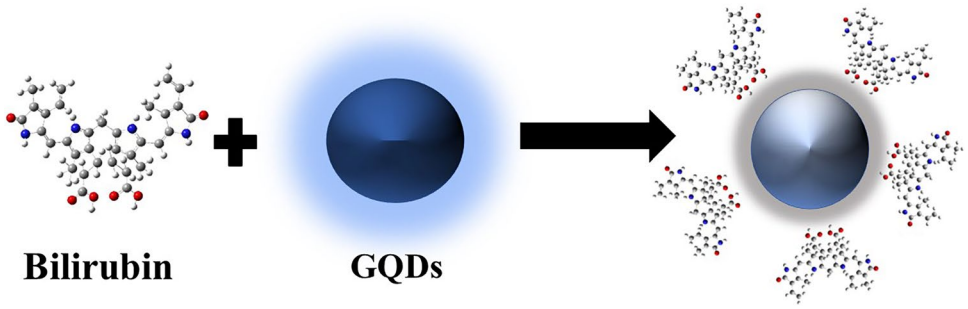


Fig. 9 Complex formation by the interaction of GQDs with BR

A linear relation indicates possibly a static quenching process involving a single kind of binding site or a possible collisional quenching mechanism. Thus, the linear SV and BH plots support predominantly the static quenching. The calculated value of K_{sv} from the S-V plot was determined to be 7.79×10^3 M^{-1}, suggesting that GQDs have significant potential for detecting bilirubin with high sensitivity. The estimated value of the limit of detection (LOD) was determined to be 87.9 nM.

As shown in Fig. 10a, the spectral overlap between the PL spectrum of GQDs as donor and the absorption spectrum of BR as acceptor indicates the chances of FRET mechanism [43]. As the PL spectrum of GQDs significantly overlaps with the absorption spectrum of BR, there is a possibility of energy transfer from excited GQDs to bilirubin molecules through FRET. The energy transfer would decrease the PL intensity of GQDs. By using an equation $J(\lambda) = \int_{\infty}^{0} F_D(\lambda)\varepsilon_A(\lambda)\lambda^4 d\lambda$, the spectral overlap integral, denoted by $J(\lambda)$, is determined. The value of $J(\lambda)$ is calculated to be $9.015 \times 10^{17} nm^4 M^{-1} cm^{-1}$ where $F_D(\lambda)$ and $\varepsilon_A(\lambda)$ represent the donor fluorescence intensity and acceptor's extinction coefficient at λ, respectively, as discussed in SI1 of supporting information (SI). Blue-emitting GQDs could detect bilirubin with the result of some parameters such as strong PL, high sensitivity, and selectivity of GQDs towards bilirubin. GQDs feature hydroxyl (–OH), epoxy, and carboxyl groups among other functional groups on their surface. These functional group interacts with the hydroxyl group of bilirubin, potentially leading to the observed quenching of GQDs PL intensity with increasing concentration of BR [43]. The PL-based sensor's ability to selectivity detect bilirubin was assessed by testing it with various biomolecules commonly encountered in biochemical engineering, including urea, uric acid, ascorbic acid, 7AI, glucose, fructose, EDTA, and BSA. Figure S1 of SI shows the absorption spectra of GQDs in the presence of BR and other biomolecules. Then, it was cleared that BR selectively quenched the PL of GQDs, whereas other biomolecules used for testing were not able to quench, as shown in Fig. 10b. These biomolecules have the potential to interfere with the specific detection of bilirubin in the biomedical field. Figure 11 shows the stability of GQDs by comparing the PL intensity as a function of months.

Conclusion

Water-dispersible GQDs with an average diameter of 3.8 ± 0.5 nm were developed from paddy straw using a hydrothermal synthesis technique. The GQDs are showing bright blue photoluminescence under 365 nm irradiation and achieve the strongest PL intensity at the 450 nm band with 370 nm excitation. The PL of GQDs was strongly quenched by the presence of bilirubin following the formation of a non-fluorescent ground-state complex (static quenching), inner filter effect (IFE), and Förster resonance energy transfer (FRET) as a quenching mechanism. The GQDs demonstrated high selectivity and sensitivity in detecting bilirubin.

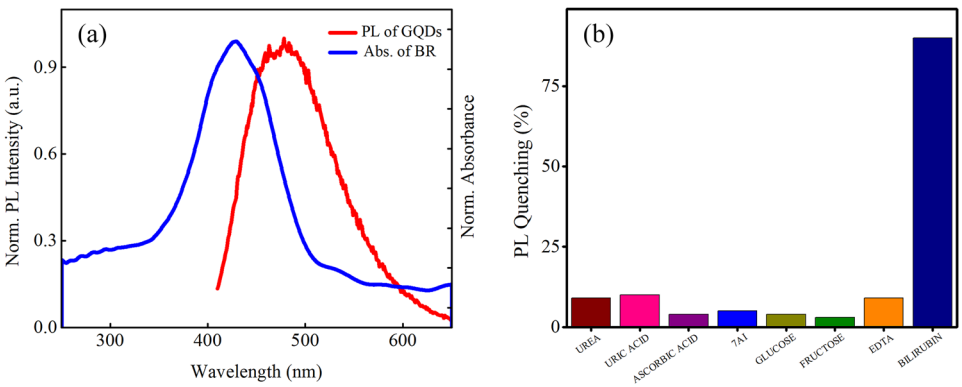


Fig. 10 Overlapped normalized absorption spectrum of BR and PL spectrum of GQDs (**a**) and a bar diagram of PL intensity of GQDs with BR and various other analogous analytes (**b**)

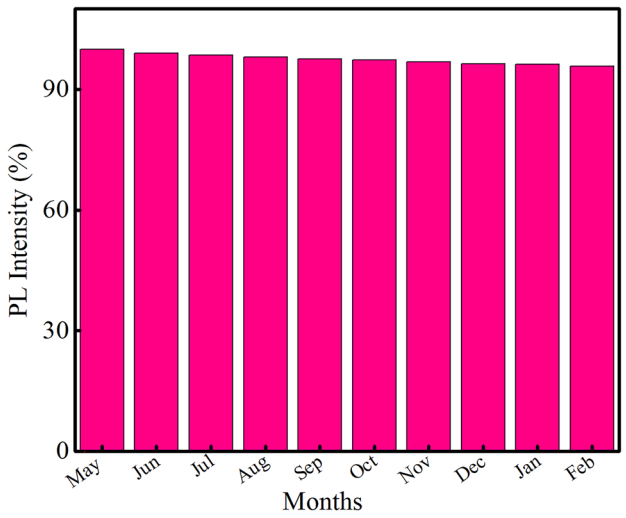


Fig. 11 PL intensity (%) of GQDs over a period of months, showing stability/photostability of GQDs

The large value of K_{sv}. i.e., 7.79×10^3 M^{-1} indicates their potential for sensitive bilirubin detection with a limit of detection (LOD) of 87.9 nM. A calibration plot for GQDs PL quenching by bilirubin showed a linear range between 9.9 and 115 µM.

Supplementary information.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11468-024-02396-0.

Author Contributions Isha: data curation, formal analysis, investigation, methodology, and writing—original draft. Aneesha: data curation and formal analysis. Mohan Singh Mehata: supervision, conceptualization, funding acquisition, resources, validation, and writing—review and editing.

Funding There is no funding for this work.

Data Availability All data generated or analyzed during this study are included in this published article.

Declarations

Competing Interests The authors declare that they have no known financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Consent to Publish All authors have approved the paper and agree with its publication.

References

1. Kumar Chaturvedi A, Pappu A, Kumar Srivastava A, Kumar Gupta M (2021) Synthesis dielectric and mechanical properties of paddy straw derived graphene quantum dots-stone waste nanocomposite. Mater Lett 301:130323
2. Duplock EJ, Scheffler M, Lindan PJD (2004) Hallmark of perfect graphene. Phys Rev Lett 92:225502
3. Wang X, Qu K, Xu B et al (2011) Microwave assisted one-step green synthesis of cell-permeable multicolor photoluminescent carbon dots without surface passivation reagents. J Mater Chem 21:2445–2450
4. Ponomarenko LA, Schedin F, Katsnelson MI et al (2008) Chaotic dirac billiard in graphene quantum dots. Science 320:356–358
5. Li LS, Yan X (2010) Colloidal graphene quantum dots. J Phys Chem Lett 1:2572–2576
6. Yuan F, Yuan T, Sui L et al (2018) Engineering triangular carbon quantum dots with unprecedented narrow bandwidth emission for multicolored LEDs. Nat Commun 9:2249
7. Ye R, Peng Z, Metzger A et al (2015) Bandgap engineering of coal-derived graphene quantum dots. ACS Appl Mater Interfaces 7:7041–7048
8. Bhatt N, Mehata MS (2023) A sustainable approach to develop gold nanoparticles with Kalanchoe fedtschenkoi and their interaction with protein and dye: sensing and catalytic probe. Plasmonics 18:845–858
9. Zhu S, Song Y, Zhao X et al (2015) The photoluminescence mechanism in carbon dots (graphene quantum dots, carbon nanodots, and polymer dots): current state and future perspective. Nano Res 8:355–381
10. Bacon M, Bradley SJ, Nann T (2014) Graphene quantum dots. Part Part Syst Charact 31:415–428
11. Zhao C, Song X, Liu Y et al (2020) Synthesis of graphene quantum dots and their applications in drug delivery. J Nanobiotechnology 18:1–32
12. Mehata MS, Biswas S (2021) Synthesis of fluorescent graphene quantum dots from graphene oxide and their application in fabrication of GQDs@AgNPs nanohybrids and sensing of H2O2. Ceram Int 47:19063–19072
13. Abbas A, Tabish TA, Bull SJ et al (2020) High yield synthesis of graphene quantum dots from biomass waste as a highly selective probe for Fe3+ sensing. Sci Rep 10:21262
14. Ray DK, Ramankutty N, Mueller ND et al (2012) Recent patterns of crop yield growth and stagnation. Nat Commun 3:1293
15. Jorn-am T, Praneerad J, Attajak R et al (2021) Quasi-solid, biorenewable supercapacitor with high specific capacitance and energy density based on rice electrolytes and rice straw-derived carbon dots as novel electrolyte additives. Colloids Surf A Physicochem Eng Asp 628:127239
16. Wang W, Martin JC, Fan X et al (2012) Silica nanoparticles and frameworks from rice husk biomass. ACS Appl Mater Interfaces 4:977–981
17. Wang Z, Yu J, Zhang X et al (2016) Large-scale and controllable synthesis of graphene quantum dots from rice husk biomass: a comprehensive utilization strategy. ACS Appl Mater Interfaces 8:1434–1439
18. Malfatti L, Innocenzi P (2018) Sol-gel chemistry for carbon dots. Chem Rec 18:1192–1202
19. Ma M, Hu X, Zhang C et al (2017) The optimum parameters to synthesize bright and stable graphene quantum dots by hydrothermal method. J Mater Sci: Mater Electron 28:6493–6497
20. He M, Guo X, Huang J et al (2018) Mass production of tunable multicolor graphene quantum dots from an energy resource of coke by a one-step electrochemical exfoliation. Carbon N Y 140:508–520
21. Calabro RL, Yang DS, Kim DY (2018) Liquid-phase laser ablation synthesis of graphene quantum dots from carbon nano-onions: comparison with chemical oxidation. J Colloid Interface Sci 527:132–140
22. Habiba K, Makarov VI, Avalos J et al (2013) Luminescent graphene quantum dots fabricated by pulsed laser synthesis. Carbon N Y 64:341–350

23. Dong LM, Shi DY, Wu Z et al (2015) Improved solvothermal method for cutting graphene oxide into graphene quantum dots. Digest Journal of Nanimaterials and Biostructures 10:855–864
24. Lu L, Zhu Y, Shi C, Pei YT (2016) Large-scale synthesis of defect-selective graphene quantum dots by ultrasonic-assisted liquid-phase exfoliation. Carbon N Y 109:373–383
25. Huang W, Zhang J, Chua SS et al (2003) Induction of bilirubin clearance by the constitutive androstane receptor (CAR). Proc Natl Acad Sci 100:4156–4161
26. Greco C, Arnolda G, Boo NY et al (2016) Neonatal jaundice in low- and middle-income countries: lessons and future directions from the 2015 Don Ostrow Trieste Yellow Retreat. Neonatology 110:172–180
27. Oza S, Lawn JE, Hogan DR et al (2015) Estimations des causes de décès néonatales pour les périodes néonatales précoces et tardives dans 194 pays: 2000–2013. Bull World Health Organ 93:19–28
28. Facure MHM, Schneider R, Mercante LA, Correa DS (2020) A review on graphene quantum dots and their nanocomposites: from laboratory synthesis towards agricultural and environmental applications. Environ Sci Nano 7:3710–3734
29. Chaturvedi AK, Pappu A, Gupta MK (2022) Unraveling the role of agro waste-derived graphene quantum dots on dielectric and mechanical property of the fly ash based polymer nanocomposite. J Alloys Compd 903:163953
30. Dubey RS, Rajesh YBRD, More MA (2015) Synthesis and characterization of SiO_2 nanoparticles via sol-gel method for industrial applications. Materials Today: Proceedings Elsevier Ltd 2:3575–3579
31. Wang W, Wang Z, Liu J et al (2018) One-pot facile synthesis of graphene quantum dots from rice husks for Fe 3+ sensing. ACS 57:1–15
32. Ibrahim SS, Elbehery HH, Samy A (2024) The efficacy of green silica nanoparticles synthesized from rice straw in the management of Callosobruchus maculatus (Col., Bruchidae). Sci Rep 14:8834
33. Abbas A, Liang Q, Abbas S et al (2022) Eco-friendly sustainable synthesis of graphene quantum dots from biowaste as a highly selective sensor. Nanomaterials 12:3696
34. Tauc J, Grigorovici R, Vancu A et al (1966) Optical properties and electronic structure of ge optical properties and electronic structure of amorphous germanium. Phys Stat Sol 15:627–637
35. He S, Turnbull MJ, Nie Y et al (2018) Band structures of blue luminescent nitrogen-doped graphene quantum dots by synchrotron-based XPS. Surf Sci 676:51–55
36. Hao YN, Guo HL, Tian L, Kang X (2015) Enhanced photoluminescence of pyrrolic-nitrogen enriched graphene quantum dots. RSC Adv 5:43750–43755
37. Li LL, Ji J, Fei R et al (2012) A facile microwave avenue to electrochemiluminescent two-color graphene quantum dots. Adv Funct Mater 22:2971–2979
38. Ananthanarayanan A, Wang X, Routh P et al (2014) Facile synthesis of graphene quantum dots from 3D graphene and their application for Fe3+ sensing. Adv Funct Mater 24:3021–3026
39. Suryawanshi A, Biswal M, Mhamane D et al (2014) Large scale synthesis of graphene quantum dots (GQDs) from waste biomass and their use as an efficient and selective photoluminescence on-off-on probe for Ag+ ions. Nanoscale 6:11664–11670
40. Croce AC, Ferrigno A, Santin G et al (2014) Bilirubin: an autofluorescence bile biomarker for liver functionality monitoring. J Biophotonics 7:810–817
41. Mehata MS (2021) An efficient excited-state proton transfer fluorescence quenching based probe (7-hydroxyquinoline) for sensing trivalent cations in aqueous environment. J Mol Liq 326:115379
42. Sharma V, Mehata MS (2021) Rapid optical sensor for recognition of explosive 2,4,6-TNP traces in water through fluorescent ZnSe quantum dots. Spectrochim Acta A Mol Biomol Spectrosc 260:119937
43. Kumar H, Obrai S (2022) Highly sensitive and selective detection of free bilirubin using blue emitting graphene quantum dots (GQDs) J Chem. Sci 134:85

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Int J Syst Assur Eng Manag
https://doi.org/10.1007/s13198-024-02371-2

ORIGINAL ARTICLE

Text mining based an automatic model for software vulnerability severity prediction

Ruchika Malhotra[1] · **Vidushi**[1]

Received: 10 January 2023 / Revised: 4 August 2023 / Accepted: 16 May 2024
© The Author(s) under exclusive licence to The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2024

Abstract Software vulnerabilities reported every year increase exponentially, leading to the exploitation of software systems. Hence, when a vulnerability is reported, a requirement arises to patch it as early as possible. Generally, this process requires some time and effort. For proper channelizing of the efforts, a requirement comes to predict the severity of the vulnerability so that the more critical ones can be given a higher priority. Therefore, a need arises to build a model that can analyze the data available on vulnerabilities and predict their severity. The experiment of this study is conducted on vulnerability reports of five software of Mozilla. As the data is textual, text mining techniques are applied to preprocess the data and form feature vectors. This input as text creates very high dimensional feature vectors leading to the requirement of dimensionality reduction. Hence, feature selection is done using chi-square and information gain. To develop the classifier, seven machine learning algorithms are chosen. Hence, fourteen software vulnerability severity prediction models (SVSPM) are developed. The result analysis allowed us to find the best-performing SVSPM. It is concluded that the model performed better for the medium and the critical severity level of the vulnerability. Out of the two feature selection techniques, information gain gave better results. An optimum number of features is also determined at which SVSPM gave good results. The best SVSPM using a machine learning algorithm corresponding to each dataset is found as well. A comparison is also made to identify significant differences among various SVSPMs developed using Friedman and Wilcoxon Signed Rank test.

Keywords Text mining · Vulnerability severity level · Feature selection · Prediction model · Machine learning

1 Introduction

A software vulnerability is defined as a weakness that attackers can exploit to intrude into a software system (Alves et al. 2016). The presence or absence of vulnerability in a software system greatly determines the quality of the software system (Erturk and Akcapinar 2015). With an increase in complexity and diversity in a software system, it is impossible to develop error-free software. Thus, taking care of the security aspect of a software system has become the utmost priority for any software developer (McGraw 2006; Furnell 2009; Violettas et al. 2013). The number of software vulnerabilities reported increases exponentially (Younan 2013; Coulter et al. 2020), and once a vulnerability is reported, its chances of getting exploited increase catastrophically (Bilge and Dumitra 2012). Therefore, developing a software vulnerability severity prediction model (SVSPM) is essential. It will automate the process of predicting the severity of software vulnerability. This helps speed up the process of prioritizing the vulnerabilities so that more severe ones can be fixed first. For the assessment of a software vulnerability, various types of vulnerability scoring frameworks (Microsoft 2022; US-CERT 2006; Sans 2022) are proposed. Among these frameworks, the common vulnerability scoring system (CVSS) (FIRST 2007) by the Forum of Incident Response and Security Teams (FIRST) is considered to be a standard framework to measure the severity of software vulnerabilities. Once a vulnerability is discovered, it is

✉ Vidushi
vidushi_2k18phdco23@dtu.ac.in

Ruchika Malhotra
ruchikamalhotra@dtu.ac.in

[1] Department of Software Engineering, Delhi Technological University, New Delhi, India

Published online: 31 May 2024

documented in the common vulnerabilities and exposures (CVE) database (Mitre 2022) which is an example of a vulnerability database. This database has vulnerability reports containing vulnerability descriptions, which are textual data.

When textual data is used as input, high-dimensional features are likely to be developed. High-dimensional data suffer from data quality problems that degrade the model's performance and increase computational cost. It is also estimated that 80–85% of the data is unstructured (Blumberg and Atre 2003; Grimes 2008). Mining such types of data is a challenge as it has various problems such as the data being unlabeled, vague, and noisy. To mitigate these problems, feature selection (FS) techniques can be applied, which help eliminate irrelevant and duplicate features and find the optimal subset that holds the maximum information, improving the performance of the model (Liu et al. 2010). Previous research in this field indicates that if FS techniques are not applied, the prediction model's performance will be degraded (Wang et al. 2019a, b; Russell et al. 2018; Gupta et al. 2015; Khazaei et al. 2016). Hence, the FS process is considered a vital aspect of the prediction process. FS techniques are broadly categorized into two techniques (1) Filter, and (2) Wrapper techniques. Researchers widely use filter techniques because of their simplicity and speed. Moreover, the result of the classifier technique does not get biased because of filter techniques. Therefore, in this study, we decided to opt for information gain (IG) and chi-square (CHI-2) FS techniques which come under filter FS techniques. Classifier is developed using machine learning (ML) algorithms. Although this idea is not new, most of the existing works are focused on security vulnerability prediction based on various derived features of the source code (Theisen and Williams 2020). As opposed to these, we propose an approach that relies on textual analysis.

The framework proposed in this study takes vulnerability reports of various software of Mozilla. These are Mozilla Firefox, Mozilla Thunderbird, Mozilla Seamonkey, Mozilla Firefox ESR, and Mozilla Thunderbird ESR. Among text mining techniques, tokenization, stop word removal, and stemming are applied to the vulnerability reports for data preprocessing. Further used bag-of-words, term frequency-inverse document frequency (TF/IDF) for the formation of feature vector and weighing terms in that feature vector, respectively. TF/IDF is a statistical technique applied most commonly in text mining (Huang et al. 2019). All these techniques convert text into numeric data. As high-dimensional feature vectors are generated, removing the curse of dimensionality becomes important. Therefore, after the formation of the feature vector, two FS filter techniques are used (IG and CHI-2) which act as fast feature selector (FFS). For the development of classifier, among ML algorithm, Bagging algorithm, Boosting algorithm, decision tree (DT), multinomial Naïve Bayes (MNB), Naïve Bayes (NB), random forest (RF), and support vector machine (SVM) are used. The selection of ML algorithms is done in such a way that we can include basic as well as more powerful ML algorithms (Sagi and Rokach 2018). Hence, fourteen models are developed using one FS technique out of two and one ML algorithm out of seven. Here, we would like to mention that the purpose of our study is not to do any comparative analysis of various text mining methodologies. We selected to work with those algorithms that are widely accepted, popularly used in the literature, and suited for our research problem.

To evaluate the performance of the SVSPM developed, AUC values are used as a performance parameter across five Mozilla datasets. The values of AUC are recorded for all four levels of severity. It is considered a good choice among various other parameters as it finds a balance between specificity and sensitivity values. Moreover, it is considered to be a nice option when dealing with imbalanced data problems (He and Garcia 2009). Statistical testing is also applied to the result values obtained to find a significant difference among various experimental scenarios, judging the effectiveness of the proposed modules for FS which is explained in detail below in this paper.

The main contribution of this study is listed as follows:

1. This paper demonstrates the significance of FS techniques, by gauging the performance of the models at different numbers of features (5, 25, 50, 100) and thus finding an optimum number of features at which the performance is maximum. Also, the effect of different levels of severity on the performance of the model is evaluated and their inference is discussed further.
2. To evaluate the effectiveness of the model developed, various experiments are conducted which help in indicating which FS technique and ML algorithm gave the best performing model.
3. Comparison is done using statistical analyses to determine if the performance of the different models differs significantly or not.

The experiments conducted in this study aimed to find the answers to the following Research Questions (RQs):

RQ1: What is the predictive capability of the SVSPM developed in this study?
AUC is used to measure the effectiveness of SVSPM. According to the values of AUC, the performance of SVSPM is analyzed.
RQ2: Does the performance of SVSPM change with a change in FS technique?
This question tries to compare the results obtained by all the SVSPM and find a conclusion about the performance of the FS technique, stating among the two which performed better.

RQ3: At which level of severity of the vulnerability, the best results are produced?
To answer this question, the results obtained are analyzed severity level-wise. The answer helps in interpreting results to conclude which level gave the highest values of the performance measures.
RQ4: Which ML algorithm works best at a particular severity level among all the ML algorithms used?
In this question, a detailed analysis of the results is done so that the name of the best ML algorithms for a particular severity level can be deduced.
RQ5: What should be the appropriate number of terms selected by the FS technique to obtain the best results?
The answer to this RQ helps in investigating the two FS techniques separately and analyzes their performance based on the selection of Top-N terms. In this study value of N is taken to be 5, 25, 50, and 100.
RQ6: Whether the performance of different SVSPMs differs significantly or not?
This RQ helps us determine the significant difference among various SVSPMs with the help of the Friedman Test and Wilcoxon Signed Rank Test.

The rest of this article is organized as follows: The related work is summarized in the next section. Section 3 lists all the preliminaries used to carry out the experiment of the paper. In Sect. 4, the proposed framework is illustrated. In Sect. 5, RQs are answered, and the results are interpreted. Threats to vulnerability are disseminated in Sect. 6. Finally, a conclusion is drawn and future work is identified in Sect. 7.

2 Related work

The key focus of this paper is on vulnerability severity prediction (i.e., predicting severity levels of vulnerabilities), rather than vulnerability prediction (i.e., classifying whether the current artifact is vulnerable or not). Therefore, in this section, previous work done in the literature matching the scope of the paper is discussed.

Han et al. (2017) implemented four baseline models and compared their performances with their proposed models. Term frequency/inverse document frequency+support vector machine (TF/IDF+SVM), word embeddings+SVM, word embedding+2-layer convolution neural network (CNN), and word embedding+CNN with long short term memory (LSTM), these four combinations are used for implementing four baseline models. The proposed approach uses word embedding+1-layer CNN combination. This combination of word embedding+1-layer CNN gave the best results with mean precision, recall, and f1-measure as 0.818, 0.815, and 0.816, respectively.

Spanos and Angelis (2018) worked on the multi-target classification of vulnerabilities. Among RF, Boosting, and DT, RF gave good results. The author specifically cited that for some vulnerability characteristics like access vector and access complexity, the results are not so good, requiring more research.

Liu et al. (2019) used CNN, LSTM, TextRCNN, and extreme gradient boosting (XGB) to develop the models. The first three technologies are the deep neural network, and the last is an ML algorithm. Results reported by the author showed that TextRCNN gave the best performance with a classification accuracy of 93.95%.

Sahin and Tosun (2019) used word embedding for feature extraction and CNN, LSTM, and XGBoost for prediction model development. Results are documented for all four levels of vulnerability, and their mean is calculated. According to the mean values, LSTM gave the best performance with a value of 0.759, 0.761, and 0.756 for precision, recall, and f1-score, respectively.

Le et al. (2019) also aimed at predicting seven vulnerability characteristics using just vulnerability descriptions. Out of seven characteristics, four are predicted well by the model developed using the light gradient boosting machine (LGBM). The model developed using XGB performed best for the other two characteristics and the logistic regression (LR) Model for the remaining one.

Wang et al. (2019a, b) used principal component analysis (PCA) on the textual description of vulnerabilities for feature extraction. SVM, LR, RF, and XGB are used for model development. Among these, the model developed using XGBoost gave the best accuracy results, which is above 90%.

The research done by Chen et al. (2020) on ten datasets having textual data on vulnerabilities. ML algorithms, namely RF, K-nearest neighbours (KNN), DT, NB, and SVM are used to build the proposed classification vulnerability model. Before applying ML algorithms, term weighing, and feature selection is done. Two techniques, TF/IDF and term frequency/inverse gravity model (TF/IGM), are used for term weighing. TF/IDF is one of the most popular and conventional methods of term weighing. On the other hand, TF/IGM is the technique proposed in this study. IG is used for FS. The result stated in the paper proves that TF/IGM is a better technique for term weighting than TF/IDF. Also, FS helps in improving the efficiency of the vulnerability classification model.

Kudjo et al. (2020) did a Bellwether analysis, which means selecting a subset of data to be treated as training data giving better performance. Deep neural networks, LR, KNN, and RF are used for model development. The results show that the Bellwether approach achieves an F-measure ranging from 14.3 to 97.8%, improving the benchmark techniques.

Babalau et al. (2021), Sharma et al. (2020), and Ni et al. (2022) also used word embedding and CNN to develop the prediction model. The experiment is conducted on different datasets, and accuracy is recorded. Ni et al. (2022) did a comparison with Sharma et al. (2020) and Han et al. (2017) and proved that his approach using bidirectional encoder representations from transformers (BERT) with CNN gave the best results.

Okutan and Mirakhorli (2022) also used CNN to build the prediction model. In this paper, the results are recorded for high and medium level of severity only.

Kekul et al. (2022) demonstrated the use of various techniques for converting text to vectors. Bag of Words (BoW), TF-IDF, 2Ngram under statistical techniques and Word2Vec, Doc2Vec, FastText under word embedding techniques. NB, DT, KNN, MLP, and RFC are used for classification.

Kudjo et al. (2023) introduced the use of the Normalized Difference Measure in this domain which was originally introduced by Rehman et al. (2017) for text classification. Five ML algorithms, namely, NB, MNB, SVM, K-NN, and DT are used for developing prediction models.

Li et al. (2023) used pre-trained language models (PLMs) such as BERT and RoBERTa as word embedding technique which is also capable of better extracting the semantic information of the text.

Mehta et al. (2023) used latent semantic indexing (LSI), latent Dirichlet allocation (LDiA), and (non-negative matrix factorization) NMF for topic modeling and LR, LDA, KNN, DT, (artificial neural network) ANN, NB, LSVM, SVM, RF, and Gaussian Naïve Bayes (GNB) for developing the classifier.

Analyzing the literature, it can be concluded that apart from ML algorithms, the use of deep learning algorithms not just for developing classifiers but also for word embeddings has given better-performing prediction models (Han et al. 2017; Sharma et al. 2020; Babalau et al. 2021; Ni et al. 2022; Kekul et al. 2022; Li et al. 2023; Mehta et al. 2023). The use of word embedding advances over traditional methods because word embedding methods also consider the semantic information of the text. Spanos and Angelis (2018) and Le et al. (2019) worked on a different objective. The prediction model developed not only predicted the severity but also predicted other features of the vulnerability report such as access vector, integrity impact, confidentiality impact, etc. Chen et al. focused on developing a new methodology for data normalization and named it TF/IGM. Unlike other studies, Şahin and Tosun (2019) and Ni et al. (2022) recorded the results for all four levels of severity of vulnerability. Wang et al. (2019a, b) recorded the performance for a different number of features to observe their influence on the performance of the prediction model.

In our study, various RQs are framed such that a comprehensive analysis can be performed on the results of the experiments conducted. The results are recorded and tabulated for different numbers of features distinctively so that the significance of FS techniques can be gauged clearly. Graphs are also framed for a better interpretation of these results. Apart from this, as results are recorded for all four levels of severity, it gave us an opportunity to do result interpretation level-wise that which level gave the best/poor performance and why. In a nutshell, it can be said that our result analysis analyzed the effect of every experimental setting on the performance of the prediction model, regarding the choice of FS technique, ML algorithm, number of features, and effect of level of severity allowing researchers to design their model carefully taking care of every experimental setting.

3 Preliminaries

This section of the paper has described the dataset, techniques, and tests used to develop the model.

3.1 Data collection

To carry out this study, we collected security vulnerabilities of six different software of Mozilla. Mozilla Firefox, Mozilla Thunderbird, Mozilla Seamonkey, Mozilla Firefox ESR, Mozilla Thunderbird ESR, and Mozilla Firefox OS from 'CVE Details,' which is a publicly available data source. After data collection, it is found that only fourteen vulnerabilities are present in Mozilla Firefox OS. As it is a very small number, therefore we have worked on the vulnerability of the remaining five datasets. The vulnerability report contained information such as ID, publish date, update date, vulnerability type, access, complexity, etc. However, we concentrated only on vulnerability description and CVSS score to conduct our experiment. Further details of the dataset are elaborated in Table 1.

CVSS is a scoring system formulated to rate vulnerabilities in the range of 0–10 (Zerkane et al. 2017). According to the most recent version of CVSS, v3.0, 'None' rating is for vulnerabilities with a 0.0 score, 'Low' severity rating is for vulnerabilities with a score of 0.1–3.9, 'Medium' rating is for vulnerabilities with a score of 4–6.9, 'High' rating is for

Table 1 Details of dataset used severity level wise

Data source	Low	Medium	High	Critical
Mozilla Firefox	164	897	296	516
Mozilla Thunderbird	22	372	164	363
Mozilla Seamonkey	13	291	61	333
Mozilla Firefox ESR	10	284	175	203
Mozilla Thunderbird ESR	1	69	15	131

vulnerabilities with a score of 7–8.9, 'Critical' rating is for vulnerabilities with a score 9–10.

3.2 Feature selection and machine learning algorithm

FS techniques are used to do dimensionality reduction. It helps reduce what is generally referred to as the 'curse' of dimensionality. This is done by selecting a subset of features and removing the features that are found to be irrelevant from the entire set. Those features which do not contribute in finding the prediction variable or the output are known as irrelevant features.

To check whether there is a dependency between feature t and category c, the researchers uses the CHI-2 (Yiming and Pederson 1997) statistical test. Ideally, it is assumed that there is no relation between feature t and category c, and both are independent of each other. If this ideal case is true, we cannot use feature t to find if the text belongs to category c. The CHI-2 test is used to describe its relevance. The value of CHI-2 is calculated according to the Eq. (1) listed below.

$$CHI\left(t_i, c_j\right) = \frac{N * (ps - qr)^2}{(p+q) * (q+r) * (r+s) * (s+p)} \tag{1}$$

In Eq. (1), N represents the total number of vulnerabilities, p represents vulnerabilities having t_i term of c_j category, q represents vulnerabilities having t_i term but not of c_j category, r represents vulnerabilities not having t_i term but of c_j category and s represents vulnerabilities neither having t_i term nor of c_j category. Hence, finally, Eq. (1) gives the CHI-2 score of term t_i and category c_j.

The term t_i is concluded to have higher significance if the chi-square value is also considerable.

According to the above stated formula, the CHI-2 score is calculated for all the terms and the top n terms having the highest score are selected after dimension reduction.

IG is the second technique used for dimensionality reduction. Menzies and Marcus (2008) in their paper found out those words that best simplify the target concept. Its concept is based on entropy. The amount of information gained from feature A given the class label as C is calculated by the following formula:

$$Infogain\left(\frac{C}{A}\right) = \mathrm{H(C)} - \mathrm{H(C/A)} \tag{2}$$

In the formula stated above, H(C) is the entropy of the class label; C and H(C/A) are the conditional entropy of C given A.

Like CHI-2, top n terms with the highest IG scores are selected after dimensionality reduction.

Among ML algorithms, the Bagging algorithm, Boosting algorithm, DT, MNB, NB, RF, and SVM are used in this study. A brief description of these algorithms can be found in Kumar and Srinivas (2021).

3.3 Statistical test selection

Hypothesis testing is done and verified using various statistical tests (Malhotra 2016). These statistical tests can be categorized as parametric and non-parametric tests. When the data samples follow specific properties, like following normal distribution, then parametric tests can be applied; else non-parametric tests will be applied. This is one way of categorizing the techniques. The other is if it is required to apply the test on two or more than two samples that are dependent or independent of each other. According to all these selection criteria, a statistical test is selected suitable for any research.

One-way ANOVA, Kruskal–Wallis, T-test, Mann Whitney U, Fridman, Paired t-test, and Wilcoxon signed-rank are some of the statistical tests. As we are dealing with more than two samples which not necessarily follow the normal distribution, the Friedman Test and Wilcoxon Signed Rank Test are found to be appropriate statistical tests for our data set.

3.4 Performance evaluation

Performance evaluation is necessary to judge the prediction power of the model developed. A confusion matrix and other parameters derived from its elements are used to do so. It is made up of four essential components, namely true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn).

Other measures for performance evaluation, such as AUC, sensitivity, specificity, accuracy, etc., can be derived using this confusion matrix and its components. One thing to underline here is that the research problem we are dealing with is a multiclass classification problem. Therefore, instead of a simple confusion matrix, the confusion matrix for multiclass looks something like depicted in Fig. 1. It shows the structure of the confusion matrix if we have three classes, C1, C2, and C3, to predict.

In this study, we used AUC as the performance measure to evaluate the effectiveness of our experiment conducted. AUC plots sensitivity values on the x-axis and specificity values on the y-axis. It is used to measure the accuracy of a model, and the value lies between 0 and 1. If the AUC value is considerable and reaches near to 1, then the prediction model is considered to have good performance. AUC as a performance measure is a popular choice among researchers as it finds a balance between both specificity and sensitivity values. Moreover, it is considered to be a good choice when dealing with imbalanced data problem (He and Garcia 2009).

Fig. 1 Confusion matrix for multiclass problem

		PREDICTED		
		C1	C2	C3
ACTUAL	C1	m1	m2	m3
	C2	m4	m5	m6
	C3	m7	m8	m9

(a) Confusion Matrix

		PREDICTED	
		C1	NOTC1
ACTUAL	C1	m1	m2+m3
	NOT C1	m4+m7	m5+m6+m8+m9

(b) Confusion Matrix for C1

		PREDICTED	
		C2	NOT C2
ACTUAL	C2	m5	m4+m6
	NOT C2	m2+m8	m1+m3+m7+m9

(c) Confusion Matrix for C2

		PREDICTED	
		C3	NOTC3
ACTUAL	C3	m9	m7+m8
	NOT C3	m3+m6	m1+m2+m4+m5

(d) Confusion Matrix for C3

4 Research methodology

The methodology followed in this study is discussed in this section. The proposed approach is depicted in Fig. 2. It involves various steps, which are elaborated below:

1. In Step 1, data is extracted, and a CSV file is formed targeting the required fields from the vulnerability reports. As five different software of Mozilla are taken into account, five different CSV files are formed. The required fields include a description of the vulnerability and the CVSS score.
2. Step 2 aims at converting the unstructured data regarding vulnerability into a structured form using various text mining techniques so that the prediction can be made further with the application of a ML algorithm.

To achieve this, data preprocessing steps are applied to the textual description of the vulnerabilities (Vijayarani et al. 2015). Tokenization, stop-word removal, and stemming are the steps involved in doing data preprocessing. Tokenization is a process of finding tokens from the text by doing its separation into words where every word is considered to be a token. After tokenization, stop word removal is done to remove all the little words from the text, allowing us to focus on essential and valuable words which contain the actual information. 'Is,' 'the,' and 'for' are some examples of stop words in the English language. In this study, we have used a default list of stop words present in python's NLTK package. Finally, stemming is performed, which is the last step of data preprocessing. Applying stemming reduces the words to their root form. For example, 'fishes', 'fishing' all are

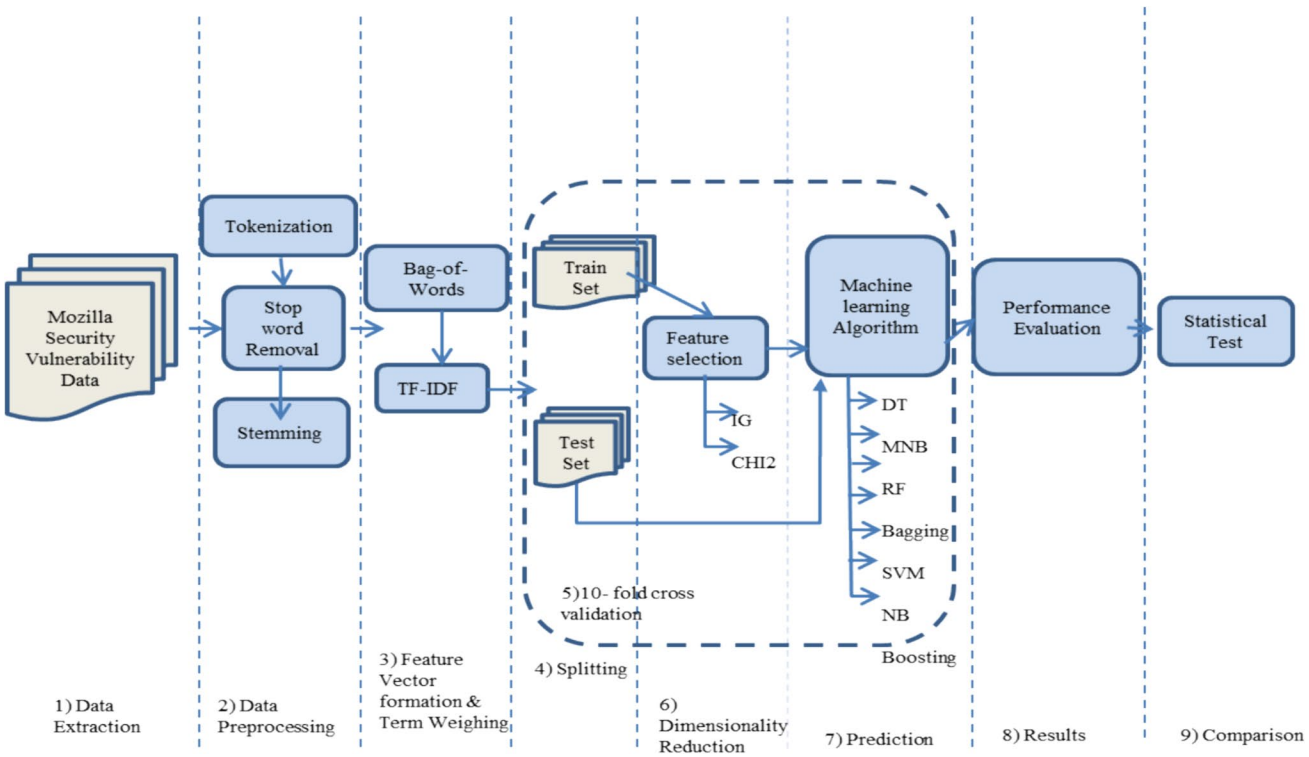


Fig. 2 Framework of our proposed approach

reduced to their root word, i.e., 'fish'. This is achieved by many researchers, and many algorithms are developed, but the most widely used and accepted algorithm is the Porter stemmer algorithm (Porter 1980). Hence, we have used it too in this study. After the application of all these three steps and completing the process of preprocessing, we are left with unique and meaningful tokens. These tokens are nothing but features that will be used by SVSPM to make predictions.

3. In Step 3, the features formed after Step 2, representing vulnerabilities, are converted into feature vectors. Bag-of-words that are formed after data preprocessing are used for the construction of these feature vectors. The length of the feature vector is the same as the number of tokens in bag-of-words. For every token in the feature vector, it contains its frequency, as seen in the textual description of the respective vulnerability. Features are reduced further by using the TF/IDF approach (Keefe and Koprinska 2009). It is a weight that is calculated for every token using TF/IDF. The calculation of the weight is done according to Eq. (3)

$$\text{For a term i in document j}: w_{i,j} = tf_{i,j} X \log\left(\frac{N}{df_i}\right) \quad (3)$$

$tf_{i,j}$ = frequency if i in j, df_i = number of descriptions containing I, N = total number of descriptions.

4. After finalizing the feature vector, the entire data is divided into training and testing data in the ratio of 80:20 as depicted in Step 4. This splitting of data is undergone under tenfold cross-validation so that much more reliable results can be obtained. In tenfold cross-validation, the entire dataset is split into 10-folds. For each ninefold of the dataset, SVSPM is developed, and performance is evaluated based on the 10th fold. This process is repeated until all the 10-folds have individually served as test data. Hence the final result is obtained by averaging the results of the ten runs of SVSPM. The training and the testing data produced using tenfold cross-validation are then fed into SVSPM. This is depicted in Step 5 of Fig. 2. Dimensionality reduction techniques are applied, which will help reduce features by selecting the most significant top-N features and removing irrelevant features. When dealing with unstructured data, the data dimension is very high, having irrelevant features. Table 2 gives the total number of textual features for individual datasets. The features that do not have any significance in finding the prediction variable or the output are considered irrelevant features. This type of dimensionality reduction is known as FS. As represented in Step 6 of Fig. 2, CHI-2 and IG are used for FS. The ML algorithms used are DT, SVM, NB, MNB, RF, Bagging, and Boosting for prediction. Hence, fourteen SVSPMs are developed, using two FS techniques and seven ML algorithms. The results are calculated using AUC values over the testing data for the performance evaluation of these prediction models.
5. While doing FS, top-5, 25, 50, and 100 terms are selected and these terms are used as attributes in the SVSPM. Therefore, when the values of AUC are recorded, the top-N terms are considered, and the results' values are recorded for this multiclass classification problem. Lastly, as can be seen in Step 9, in addition to performance measurement, hypothesis testing is done using the Friedman test and Wilcoxon Signed Rank test. Null and alternative hypotheses are formed (Ho and Ha) and are accessed at a 0.05 level of significance. This test helps check whether the difference in the effectiveness measure of the ML algorithms is statistically significant or not. Therefore, the hypothesis is stated as below:

   Ho = The performance of all SVSPM developed using different ML algorithms is the same.
   Ha = There is a significant difference between the performance of all SVSPM developed using various ML algorithms.

Based on the rejection and acceptance of these hypotheses, the Friedman test will help rank all the SVSPMs, whereas the Wilcoxon Signed Rank test will help do a pairwise comparison of SVSPMs.

5 Result analysis

In this section, we will answer the RQs which are discussed above. After analyzing the result concerning a particular RQ, it is concluded with a meaningful finding.

Table 2 Size of Dataset in terms of number of vulnerabilities and word count

	Mozilla Firefox	Mozilla Thunderbird	Mozilla Seamonkey	Mozilla Firefox ESR	Mozilla Thunderbird ESR
Vulnerabilities	1873	921	698	672	216
Text features	23,887	11,903	8862	9094	2841

RQ1: What is the predictive capability of the SVSPM developed in this study?

Tables 3, 4, 5, 6 and 7 present the performance of the models using AUC over the FS techniques with respect to the top 5, 25, 50, and 100 terms. For Mozilla Thunderbird ESR, the value of AUC for low severity level of vulnerability is 'nan' as only one vulnerability is present, which is depicted in Table 1. For a detailed analysis of the effect of the FS technique, the values of AUC are recorded at four different values of N, i.e., 5, 25, 50, and 100. After analyzing the results, the following conclusion is obtained:

1. For all five datasets, irrespective of the FS technique and ML algorithm used, we can conclude that better results are obtained for the top-100 terms.
2. Secondly, result values are also varying as the level of severity of the vulnerability changes. A typical pattern is observed for all the datasets. Irrespective of the FS technique and ML algorithm, better results are obtained for the medium and the critical level of severity of vulnerabilities.

Table 8 is constructed in which the fourteen models which are developed in this study are given names according to the FS technique and ML algorithm used. These model names are used below in the text while addressing RQs.

Table 3 Results obtained for Mozilla Firefox

AUC values using chi-square						AUC values using info gain					
Top-5 terms						Top-5 terms					
Dataset	Algorithm	Low	Medium	High	Critical	Dataset	Algorithm	Low	Medium	High	Critical
Mozilla Firefox	DT	0.5	0.49	0.49	0.62	Mozilla Firefox	DT	0.44	0.52	0.46	0.71
	SVM	0.5	0.5	0.5	0.61		SVM	0.5	0.5	0.5	0.61
	NB	0.5	0.6	0.6	0.63		NB	0.51	0.73	0.41	0.69
	Bagging	0.5	0.49	0.5	0.62		Bagging	0.48	0.68	0.52	0.7
	MNB	0.5	0.5	0.5	0.5		MNB	0.49	0.49	0.45	0.55
	RF	0.5	0.49	0.49	0.62		RF	0.49	0.66	0.48	0.69
	Boosting	0.5	0.5	0.49	0.7		Boosting	0.5	0.57	0.44	0.7
Top-25 terms						Top-25 terms					
Mozilla Firefox	DT	0.5	0.49	0.5	0.74	Mozilla Firefox	DT	0.49	0.66	0.59	0.74
	SVM	0.5	0.5	0.5	0.61		SVM	0.5	0.52	0.5	0.68
	NB	0.5	0.63	0.59	0.71		NB	0.51	0.7	0.66	0.75
	Bagging	0.5	0.55	0.44	0.73		Bagging	0.48	0.7	0.52	0.72
	MNB	0.5	0.5	0.5	0.61		MNB	0.5	0.5	0.49	0.61
	RF	0.5	0.7	0.5	0.69		RF	0.49	0.7	0.48	0.69
	Boosting	0.5	0.67	0.44	0.71		Boosting	0.5	0.57	0.49	0.74
Top-50 terms						Top-50 terms					
Mozilla Firefox	DT	0.47	0.6	0.51	0.75	Mozilla Firefox	DT	0.79	0.63	0.67	0.77
	SVM	0.5	0.5	0.5	0.64		SVM	0.5	0.56	0.5	0.68
	NB	0.54	0.63	0.6	0.71		NB	0.55	0.67	0.54	0.67
	Bagging	0.65	0.55	0.48	0.74		Bagging	0.51	0.7	0.55	0.72
	MNB	0.5	0.5	0.5	0.61		MNB	0.5	0.5	0.49	0.63
	RF	0.65	0.54	0.46	0.69		RF	0.5	0.67	0.52	0.72
	Boosting	0.5	0.6	0.47	0.75		Boosting	0.5	0.56	0.52	0.73
Top-100 terms						Top-100 terms					
Mozilla Firefox	DT	0.63	0.67	0.59	0.77	Mozilla Firefox	DT	0.47	0.63	0.53	0.71
	SVM	0.5	0.55	0.5	0.67		SVM	0.5	0.56	0.49	0.7
	NB	0.6	0.6	0.61	0.7		NB	0.61	0.71	0.78	0.66
	Bagging	0.47	0.53	0.54	0.74		Bagging	0.61	0.7	0.57	0.75
	MNB	0.5	0.53	0.5	0.64		MNB	0.5	0.51	0.49	0.64
	RF	0.5	0.57	0.61	0.69		RF	0.5	0.7	0.47	0.73
	Boosting	0.5	0.67	0.48	0.75		Boosting	0.5	0.7	0.53	0.74

Table 4 Result obtained for Mozilla Thunderbird

AUC values using chi-square						AUC values using info gain					
Top-5 terms						Top-5 terms					
Dataset	Algorithm	Low	Medium	High	Critical	Dataset	Algorithm	Low	Medium	High	Critical
Mozilla Thunder	DT	0.5	0.49	0.49	0.62	Mozilla Thunderbird	DT	0.49	0.66	0.45	0.71
	SVM	0.5	0.5	0.5	0.5		SVM	0.44	0.52	0.59	0.74
	NB	0.5	0.5	0.5	0.5		NB	0.6	0.69	0.65	0.63
	Bagging	0.5	0.5	0.5	0.5		Bagging	0.45	0.58	0.53	0.69
	MNB	0.5	0.5	0.5	0.48		MNB	0.49	0.6	0.53	0.7
	RF	0.5	0.5	0.5	0.5		RF	0.47	0.55	0.47	0.65
	Boosting	0.5	0.5	0.5	0.5		Boosting	0.5	0.56	0.5	0.67
Top-25 terms						Top-25 terms					
Mozilla Thunder	DT	0.5	0.59	0.5	0.7	Mozilla Thunder	DT	0.45	0.58	0.55	0.69
	SVM	0.5	0.61	0.5	0.69		SVM	0.45	0.51	0.49	0.65
	NB	0.56	0.5	0.44	0.47		NB	0.58	0.67	0.63	0.63
	Bagging	0.5	0.55	0.49	0.56		Bagging	0.43	0.58	0.53	0.65
	MNB	0.5	0.43	0.5	0.57		MNB	0.44	0.6	0.53	0.7
	RF	0.5	0.54	0.5	0.59		RF	0.48	0.55	0.47	0.65
	Boosting	0.5	0.5	0.5	0.5		Boosting	0.5	0.56	0.5	0.71
Top-50 terms						Top-50 terms					
Mozilla Thunder	DT	0.5	0.62	0.5	0.72	Mozilla Thunder	DT	0.45	0.58	0.55	0.69
	SVM	0.47	0.52	0.55	0.74		SVM	0.47	0.53	0.49	0.66
	NB	0.57	0.55	0.53	0.52		NB	0.6	0.69	0.65	0.63
	Bagging	0.53	0.57	0.51	0.57		Bagging	0.45	0.58	0.53	0.69
	MNB	0.54	0.45	0.49	0.58		MNB	0.49	0.6	0.53	0.7
	RF	0.5	0.55	0.5	0.63		RF	0.47	0.55	0.47	0.65
	Boosting	0.5	0.49	0.5	0.63		Boosting	0.5	0.57	0.5	0.69
Top-100 terms						Top-100 terms					
Mozilla Thunder	DT	0.5	0.67	0.5	0.74	Mozilla Thunder	DT	0.47	0.6	0.55	0.71
	SVM	0.63	0.49	0.5	0.75		SVM	0.5	0.55	0.49	0.74
	NB	0.48	0.5	0.52	0.7		NB	0.62	0.71	0.68	0.66
	Bagging	0.5	0.55	0.51	0.69		Bagging	0.47	0.6	0.55	0.71
	MNB	0.5	0.48	0.49	0.63		MNB	0.5	0.61	0.55	0.72
	RF	0.5	0.58	0.5	0.69		RF	0.5	0.57	0.5	0.69
	Boosting	0.63	0.64	0.5	0.75		Boosting	0.5	0.6	0.5	0.72

RQ2: Does the performance of SVSPM change with the change in FS technique?

As analyzed by RQ1, the best results are obtained for the top 100 terms. Therefore, to answer this RQ, values recorded for top-100 terms in Tables 3, 4, 5, 6 and 7 are analyzed. Analyzing these values, Table 9 is framed which contains the best and the second-best AUC values for all five databases. Apart from AUC values, the name of the model, the technology used by the model, and the corresponding severity level are also recorded. Looking at these observations of Table 9, it is observed that out of the top fifteen AUC values, nine AUC values are recorded when IG is used as the FS technique. Also, for all five datasets, if not the best, the second-best result is obtained when the IG FS technique is used. Graph of Fig. 3 is also constructed which plots the mean AUC values over five datasets with respect to top-100 terms for CHI-2 and IG FS techniques. Looking at this graph, we can say that IG produces better results if compared with CHI-2. Hence, we can conclude that when the AUC values are observed for an individual model or when the mean AUC values are observed, in both cases, IG came out to be a better-performing FS technique as compared to CHI-2.

RQ3: At which level of severity of the vulnerability, the best results are produced?

Table 5 Result obtained for Mozilla Seamonkey

AUC values using chi-square						AUC values using info gain					
Top-5 terms						Top-5 terms					
Dataset	Algorithm	Low	Medium	High	Critical	Dataset	Algorithm	Low	Medium	High	Critical
Mozilla Sea	DT	0.5	0.5	0.5	0.5	Mozilla Sea	DT	0.49	0.66	0.45	0.71
	SVM	0.5	0.5	0.5	0.5		SVM	0.44	0.52	0.59	0.74
	NB	0.5	0.5	0.5	0.5		NB	0.59	0.65	0.59	0.72
	Bagging	0.5	0.5	0.5	0.5		Bagging	0.5	0.71	0.54	0.71
	MNB	0.5	0.5	0.5	0.5		MNB	0.5	0.59	0.72	0.57
	RF	0.5	0.5	0.5	0.5		RF	0.5	0.72	0.5	0.8
	Boosting	0.5	0.5	0.5	0.5		Boosting	0.5	0.74	0.5	0.77
Top-25 terms						Top-25 terms					
Mozilla Sea	DT	0.5	0.71	0.51	0.65	Mozilla Sea	DT	0.44	0.52	0.59	0.7
	SVM	0.5	0.69	0.5	0.74		SVM	0.55	0.7	0.45	0.75
	NB	0.6	0.51	0.53	0.52		NB	0.59	0.65	0.59	0.72
	Bagging	0.5	0.71	0.5	0.62		Bagging	0.5	0.71	0.54	0.71
	MNB	0.5	0.38	0.5	0.61		MNB	0.5	0.59	0.72	0.57
	RF	0.5	0.69	0.51	0.66		RF	0.5	0.72	0.5	0.8
	Boosting	0.5	0.74	0.51	0.73		Boosting	0.5	0.75	0.5	0.82
Top-50 terms						Top-50 terms					
Mozilla Sea	DT	0.5	0.68	0.59	0.72	Mozilla Sea	DT	0.79	0.63	0.67	0.77
	SVM	0.5	0.69	0.5	0.74		SVM	0.5	0.7	0.5	0.7
	NB	0.6	0.53	0.6	0.58		NB	0.59	0.65	0.59	0.72
	Bagging	0.5	0.77	0.5	0.73		Bagging	0.5	0.71	0.54	0.71
	MNB	0.5	0.47	0.5	0.7		MNB	0.5	0.59	0.72	0.57
	RF	0.5	0.7	0.48	0.78		RF	0.5	0.72	0.5	0.8
	Boosting	0.5	0.75	0.5	0.77		Boosting	0.5	0.74	0.49	0.82
Top-100 terms						Top-100 terms					
Mozilla Sea	DT	0.5	0.67	0.52	0.72	Mozilla Sea	DT	0.5	0.66	0.43	0.67
	SVM	0.5	0.73	0.5	0.81		SVM	0.5	0.75	0.5	0.77
	NB	0.5	0.54	0.59	0.64		NB	0.59	0.65	0.67	0.72
	Bagging	0.5	0.68	0.59	0.7		Bagging	0.5	0.71	0.54	0.71
	MNB	0.5	0.49	0.6	0.74		MNB	0.5	0.59	0.72	0.57
	RF	0.5	0.71	0.5	0.81		RF	0.5	0.72	0.5	0.8
	Boosting	0.5	0.71	0.49	0.82		Boosting	0.5	0.74	0.5	0.82

In this RQ, the AUC values recorded in Tables 3, 4, 5, 6 and 7 are analyzed in such a way that it can be figured out whether the performance of the model is dependent on the level of severity of the vulnerability. For this purpose, graphs of Fig. 4 are constructed which plots the performance of the models developed for all five databases FS technique-wise. A common pattern is observed in these graphs that the AUC values obtained for the medium and the critical level of severity of vulnerability are much higher as compared with the AUC values of the low and the high level of severity of the vulnerability.

Hence, it can be concluded that irrespective of FS technique and ML algorithms, SVSPM gave better performance better for the critical and the medium level of severity as compared to the low and the high level of severity.

RQ4: Which ML algorithm works best at a particular severity level among all the ML algorithms used?

This RQ tries to find the best among all the ML algorithms which are used to conduct the experiment of the study, level-wise. Keeping in mind the conclusion drawn from the above RQs, we concentrated on the values produced by selecting top-100 terms. To obtain the best algorithm for all five datasets at a particular severity level, the algorithm giving the best and second-best value of AUC is recorded in Table 10. The second-best AUC value is recorded only if the

Table 6 Result obtained for Mozilla Firefox ESR

AUC values using chi-square						AUC values using info gain					
Top-5 terms						Top-5 terms					
Dataset	Algorithm	Low	Medium	High	Critical	Dataset	Algorithm	Low	Medium	High	Critical
Mozilla Firefox ESR	DT	0.5	0.55	0.5	0.56	Mozilla Firefox ESR	DT	0.5	0.82	0.5	0.72
	SVM	0.5	0.5	0.5	0.5		SVM	0.5	0.88	0.5	0.7
	NB	0.5	0.5	0.5	0.5		NB	0.5	0.83	0.5	0.77
	Bagging	0.5	0.5	0.5	0.5		Bagging	0.5	0.8	0.58	0.72
	MNB	0.5	0.5	0.5	0.5		MNB	0.5	0.72	0.5	0.8
	RF	0.5	0.5	0.5	0.5		RF	0.5	0.81	0.5	0.77
	Boosting	0.5	0.73	0.5	0.54		Boosting	0.49	0.82	0.48	0.75
Top-25 terms						Top-25 terms					
Mozilla Firefox ESR	DT	0.5	0.59	0.5	0.63	Mozilla Firefox ESR	DT	0.5	0.85	0.5	0.79
	SVM	0.5	0.7	0.5	0.75		SVM	0.5	0.9	0.5	0.8
	NB	0.6	0.51	0.53	0.52		NB	0.5	0.88	0.5	0.8
	Bagging	0.5	0.72	0.51	0.66		Bagging	0.5	0.8	0.58	0.75
	MNB	0.5	0.39	0.5	0.62		MNB	0.5	0.77	0.5	0.81
	RF	0.5	0.69	0.5	0.66		RF	0.5	0.85	0.5	0.8
	Boosting	0.5	0.61	0.5	0.6		Boosting	0.5	0.91	0.5	0.84
Top-50 terms						Top-50 terms					
Mozilla Firefox ESR	DT	0.5	0.85	0.49	0.73	Mozilla Firefox ESR	DT	0.5	0.87	0.5	0.8
	SVM	0.5	0.77	0.5	0.8		SVM	0.5	0.93	0.5	0.82
	NB	0.6	0.53	0.6	0.57		NB	0.5	0.9	0.5	0.81
	Bagging	0.5	0.76	0.51	0.73		Bagging	0.5	0.82	0.55	0.77
	MNB	0.5	0.47	0.5	0.7		MNB	0.5	0.79	0.5	0.84
	RF	0.5	0.71	0.48	0.78		RF	0.5	0.88	0.5	0.81
	Boosting	0.5	0.92	0.5	0.84		Boosting	0.5	0.94	0.5	0.82
Top-100 terms						Top-100 terms					
Mozilla Firefox ESR	DT	0.5	0.85	0.4	0.73	Mozilla Firefox ESR	DT	0.5	0.91	0.5	0.82
	SVM	0.5	0.74	0.5	0.81		SVM	0.5	0.97	0.5	0.85
	NB	0.5	0.54	0.58	0.64		NB	0.5	0.97	0.5	0.85
	Bagging	0.5	0.71	0.59	0.76		Bagging	0.5	0.87	0.58	0.8
	MNB	0.5	0.49	0.6	0.73		MNB	0.5	0.85	0.5	0.85
	RF	0.5	0.71	0.5	0.81		RF	0.5	0.93	0.5	0.84
	Boosting	0.5	0.92	0.5	0.84		Boosting	0.5	0.92	0.5	0.84

difference between the two is not more than 0.05. Grey cells in Table 10 represent that the documentation of the second-best value is not needed and ML algorithm name which is in bold is the one giving good performance for all datasets. Looking at the values recorded in Table 10, for the low level of severity, no ML algorithm is able to produce good results. For the medium level of severity, if not the best, the second-best value is achieved by Boosting algorithm. Similarly, for the high level of severity, the common ML algorithm which gave good results is the NB ML algorithm. For the critical level of severity, the common ML algorithm which gave good results is the Boosting ML algorithm.

From the above analysis, we come to the conclusion that, out of four, boosting gave good results for two levels of severity. Hence, we can say that, apart from traditional ML algorithms, more powerful ensemble boosting technique gave promising results.

RQ5. What should be the appropriate number of terms selected by the FS techniques to obtain the best results?

The number of terms is the value of N in top-N terms. These are the most significant terms having the most importance out of all text features. A very small value of N will lead to a loss of information, and subsequently, the efficiency of the model will be low. In contrast, a very big value of N will lead to an increase in dimensionality, leading to an increase in the complexity of the model. Therefore, a search for the optimum value of N is required. Keeping this in mind, we executed our experiment for four different values

Table 7 Result obtained for Mozilla Thunderbird ESR

AUC values using chi-square						AUC values using info gain					
Top-5 terms						Top-5 terms					
Dataset	Algorithm	Low	Medium	High	Critical	Dataset	Algorithm	Low	Medium	High	Critical
Mozilla Thunder	DT	Nan	0.5	0.5	0.5	Mozilla Thunder	DT	Nan	0.87	0.5	0.7
	SVM	Nan	0.5	0.5	0.5		SVM	Nan	0.8	0.5	0.73
	NB	Nan	0.5	0.5	0.5		NB	Nan	0.79	0.5	0.68
	Bagging	Nan	0.5	0.5	0.5		Bagging	Nan	0.88	0.5	0.7
	MNB	Nan	0.5	0.5	0.5		MNB	Nan	0.66	0.5	0.65
	RF	Nan	0.5	0.5	0.5		RF	Nan	0.87	0.5	0.71
	Boosting	Nan	0.5	0.5	0.5		Boosting	Nan	1	0.5	0.81
Top-25 terms						Top-25 terms					
Mozilla Thunde	DT	Nan	0.8	0.5	0.685	Mozilla Thunder	DT	Nan	0.9	0.5	0.7
	SVM	Nan	0.5	0.5	0.5		SVM	Nan	0.85	0.5	0.75
	NB	Nan	0.45	0.62	0.56		NB	Nan	0.8	0.5	0.7
	Bagging	Nan	0.8	0.5	0.69		Bagging	Nan	0.91	0.5	0.72
	MNB	Nan	0.5	0.5	0.5		MNB	Nan	0.7	0.5	0.66
	RF	Nan	0.8	0.5	0.68		RF	Nan	0.9	0.5	0.73
	Boosting	Nan	0.7	0.5	0.62		Boosting	Nan	0.95	0.5	0.84
Top-50 terms						Top-50 terms					
Mozilla Thunder	DT	Nan	0.95	0.5	0.81	Mozilla Thunder	DT	Nan	0.9	0.5	0.7
	SVM	Nan	0.75	0.5	0.75		SVM	Nan	0.87	0.5	0.78
	NB	Nan	0.49	0.72	0.68		NB	Nan	0.8	0.5	0.71
	Bagging	Nan	0.95	0.5	0.81		Bagging	Nan	0.94	0.5	0.77
	MNB	Nan	0.6	0.5	0.56		MNB	Nan	0.74	0.5	0.69
	RF	Nan	0.95	0.5	0.78		RF	Nan	0.91	0.5	0.76
	Boosting	Nan	1	0.5	0.81		Boosting	Nan	1	0.5	0.78
Top-100 terms						Top-100 terms					
Mozilla Thunder	DT	Nan	1	0.5	0.81	Mozilla Thunder	DT	Nan	0.95	0.5	0.72
	SVM	Nan	0.8	0.5	0.72		SVM	Nan	0.9	0.5	0.78
	NB	nan	0.53	0.72	0.72		NB	Nan	0.85	0.5	0.71
	Bagging	Nan	1	0.5	0.78		Bagging	Nan	0.95	0.5	0.78
	MNB	Nan	0.6	0.5	0.652		MNB	Nan	0.8	0.5	0.69
	RF	Nan	0.95	0.5	0.78		RF	Nan	0.95	0.5	0.78
	Boosting	Nan	1	0.5	0.81		Boosting	Nan	1	0.5	0.81

Table 8 Models developed

Model	FS technique	ML algorithm	Model	FS technique	ML algorithm
Model 1	CHI-2	DT	Model 8	IG	DT
Model 2	CHI-2	SVM	Model 9	IG	SVM
Model 3	CHI-2	NB	Model 10	IG	NB
Model 4	CHI-2	Bagging	Model 11	IG	Bagging
Model 5	CHI-2	MNB	Model 12	IG	MNB
Model 6	CHI-2	RF	Model 13	IG	RF
Model 7	CHI-2	Boosting	Model 14	IG	Boosting

Table 9 Best and Second-Best AUC Values for CHI-2 and IG FS technique

AUC value		Model	Technology used	Severity level	Database
0.78	Best	Model 10	IG + NB	High	Mozilla Firefox
0.77	Second Best	Model 8	IG + DT	Critical	Mozilla Firefox
0.75	Best	Model 2, Model 7	CHI-2 + SVM, CHI-2 + Boosting	Critical	Mozilla Thunderbird
0.74	Second Best	Model 1, Model 9	CHI-2 + DT, IG + SVM	Critical	Mozilla Thunderbird
0.82	Best	Model 7, Model 14	CHI-2 + Boosting, IG + Boosting	Critical	Mozilla Seamonkey
0.81	Second Best	Model 2, Model 6	CHI-2 + SVM, CHI-2 + RF	Critical	Mozilla Seamonkey
0.97	Best	Model 10, Model 9	IG + NB, IG + SVM	Medium	Mozilla Firefox ESR
0.93	Second Best	Model 13	IG + RF	Medium	Mozilla Firefox ESR
0.95	Best	Model 8	IG + DT	Medium	Mozilla Thunderbird ESR
0.93	Second Best	Model 9	IG + SVM	Medium	Mozilla Thunderbird ESR

of N, which are 5, 25, 50, and 100. The results are recorded in Tables 3, 4, 5, 6 and 7 which contain the AUC values of all these four instances. Using these results as the base, we did its analysis and tried to obtain if any relation between the AUC measure and the value of N exists.

We plotted a graph between the number of terms, i.e., N, and corresponding AUC values to do the analysis. There are many different combinations available for graph plotting because we had four levels of severity, two FS techniques, seven ML algorithms, and five datasets. For our simplicity, we considered the level of severity that gave the best results, which is the critical level, as discussed in RQ3. Among all ML algorithms, we considered the algorithm that gave the best results at the critical level, which is boosting, as discussed in RQ4. The graphs are depicted in Fig. 5 in which the x-axis represents the number of features, and the y-axis represents the value of AUC.

Looking at the lines in the graph, we can conclude that as the value of N or the number of features increases, the performance of SVSPM gets better and later on becomes stable or slightly reduced due to the addition of some redundant or irrelevant features.

RQ6. Whether the performance of all the SVSPM developed differ significantly or not?

To check whether the performance of all the ML algorithms differs significantly or not, the Friedman test is used. The hypothesis is framed and is stated as given below:

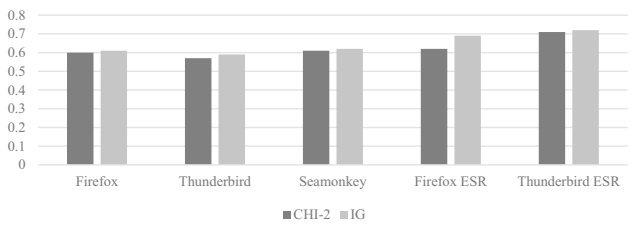


Fig. 3 Plot of Mean AUC values database-wise for CHI-2 and IG FS Techniques

Ho = The performance of all SVSPM developed using different ML algorithms is the same.
Ha = There is a significant difference between the performance of all SVSPM developed using various ML algorithms.

The input to the Friedman test is taken from Tables 3, 4, 5, 6 and 7. From the available set of AUC values, we have taken values as input to the Friedman test that is recorded for the top 100 terms. For low level, 0.692 and 0.066 are the p-values that came when CHI-2 and IG are used for FS respectively. For the medium level, 0.004 and 0.286 are the *p* values that came when CHI-2 and IG are used for FS respectively. For the high level, 0.053 and 0.054 are the *p* values that came when CHI-2 and IG are used for FS respectively. For the critical level, 0.014 and 0.166 are the *p* values that came when CHI-2 and IG are used for FS respectively. The results of the Friedman test showing the mean rank for all SVSPM are documented in Table 11.

According to the *p* values recorded, Ho is retained for all four levels of severity, when IG is used as FS technique. This depicts that there is no significant difference in the models' performance. Hence, Friedman's test results also validate the finding of RQ2 that is the IG FS technique gives better results. Therefore, when the IG FS technique is used, all SVSPM gave a good performance with no significant difference in result values. Ho is retained for low and medium level of severity as well. This friedman's test result validates the finding of RQ3. Analysis of RQ3 tells that SVSPM performance was not good for low and high level of severity. Hence as bad results came for both the levels for all ML algorithms, therefore no significant difference exists in the performance of SVSPM.

Looking at the *p* values for the medium and critical level of severity, Ho is rejected when CHI-2 is used as a FS technique.

Wilcoxon signed-rank test is used to extend our investigations pairwise for the above-stated cases in which Ho is rejected. The pairwise analysis will help us determine which

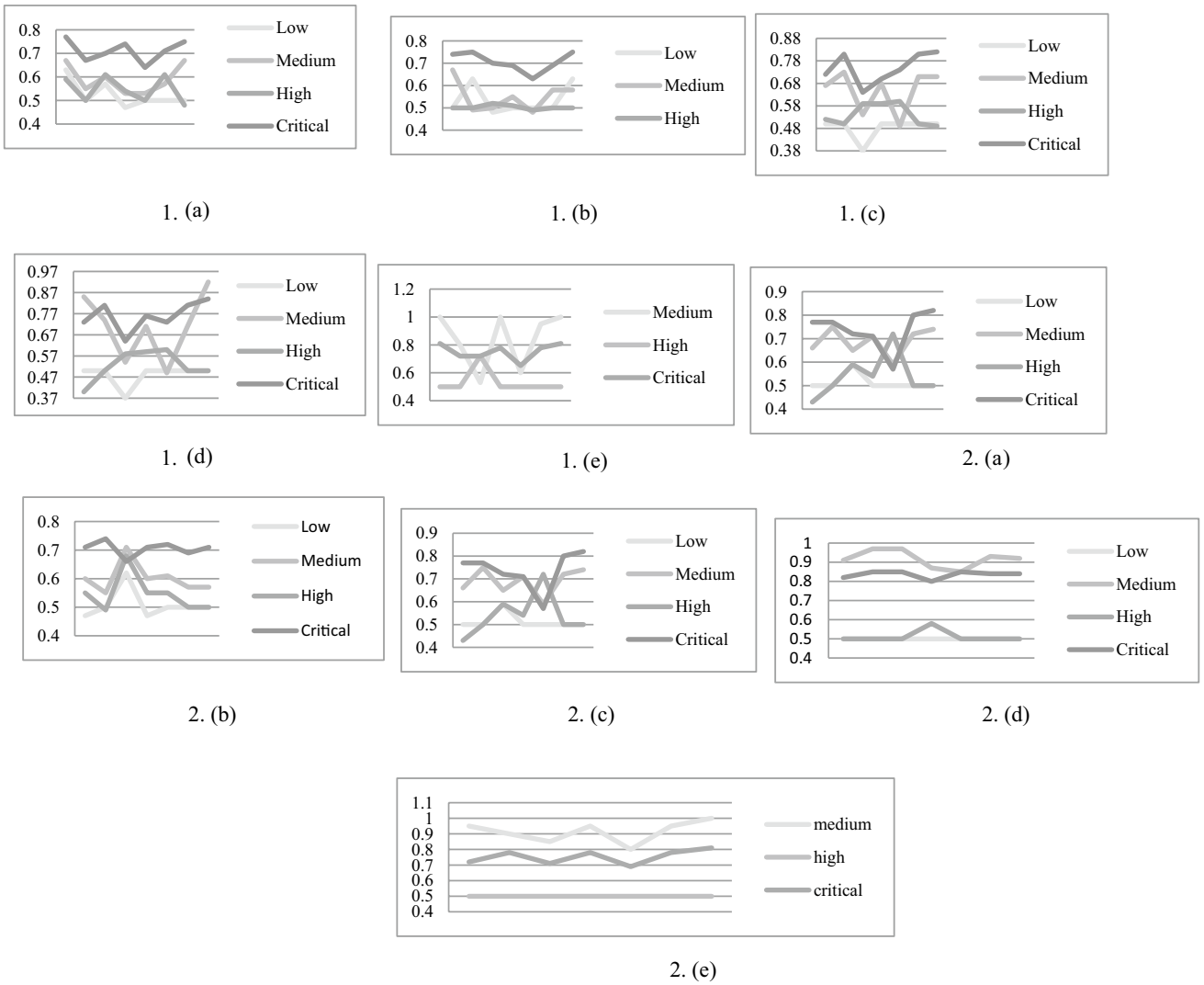


Fig. 4 The AUC values obtained severity level wise for all datasets with 1. **a** CHI-2 for Mozilla Firefox, **b** CHI-2 for Mozilla Thunderbird, **c** CHI-2 for Mozilla Seamonkey, **d** CHI-2 for Mozilla Firefox ESR, **e** CHI-2 for Mozilla Thunderbird ESR and 2. **a** IG for Mozilla Firefox, **b** IG for Mozilla Thunderbird, **c** IG for Mozilla Seamonkey, **d** IG for Mozilla Firefox ESR, **e** IG for Mozilla Thunderbird ESR

pair of ML algorithms significantly differ from each other. Hence, for Wilcoxon Signed rank test, the hypothesis framed are stated as follows:

- Ho: No significant difference is there between the performance of two SVSPM
- Ha: Significant difference occurs between the performance of two SVSPM

The results are tabulated in Table 12. The values in bold in Table 12 depicts that, for that particular pair of ML algorithm, a significant difference exists in the result values of the model. If we focus on Boosting algorithm, a total of six pairs are made with it. Out of six pairs, the following four Boosting-NB, Boosting-Bagging, Boosting-MNB, and Boosting-RF showed significant differences in performance. Hence a conclusion can be drawn that SVSPM developed using boosting algorithm are having good performance and gave good results for the medium and critical level of severity. This validated our findings of an analysis of RQ4 as well. Apart from these, the other pairs which showed significant differences are NB-DT, MNB-DT, and MNB- SVM for medium level and NB-DT, MNB-SVM for critical level.

6 Threats to validity

The possible validity threats to our study are as follows:

Table 10 AUC values of best and second best performing ML algorithms database wise for all Levels of severity of Vulnerability

		Mozilla Firefox		Mozilla Thunderbird		Mozilla Seamonkey		Mozilla Firefox ESR		Mozilla Thunderbird ESR	
		AUC value	ML algorithm	AUC value	ML algorithm	AUC value	ML algorithm	AUC value	ML algorithm	AUC value	ML algorithm
CHI-2	**Low**	0.63	DT	0.63	SVM, Boosting	0.5	All	0.5	All	–	–
		0.61	NB								
	Medium	0.67	DT, **Boosting**	0.67	DT	0.73	SVM	0.92	**Boosting**	1	**Boosting**
				0.64	**Boosting**	0.71	RF, **Boosting**				
	High	0.61	**NB**	0.52	**NB**	0.6	MNB	0.59	Bagging	0.72	**NB**
						0.59	**NB**	0.58	**NB**, Boosting		
	Critical	0.77	DT	0.75	SVM, **Boosting**	0.82	**Boosting**	0.84	**Boosting**	0.81	**Boosting,** DT
		0.75	**Boosting**								
IG	**Low**	0.61	**NB,** Bagging	0.62	**NB**	0.59	**NB**	0.5	All	–	–
	Medium	0.71	NB	0.61	MNB	0.75	SVM	0.97	NB	1	**Boosting**
		0.7	Bagging, RF, **Boosting**	0.6	DT, Bagging, **Boosting**	0.74	**Boosting**	0.93, 0.92	RF, **Boosting**		
	High	0.78	**NB**	0.68	**NB**	0.72	MNB	0.5	All	0.5	All
						0.67	**NB**				
	Critical	0.75	Bagging	0.74	SVM	0.82	**Boosting**	0.85	SVM, NB, MNB	0.81	**Boosting**
		0.74	**Boosting**	0.72	**Boosting**						

The ML algorithm giving the best result for all datasets, at a particular level of severity of vulnerability are shown in bold

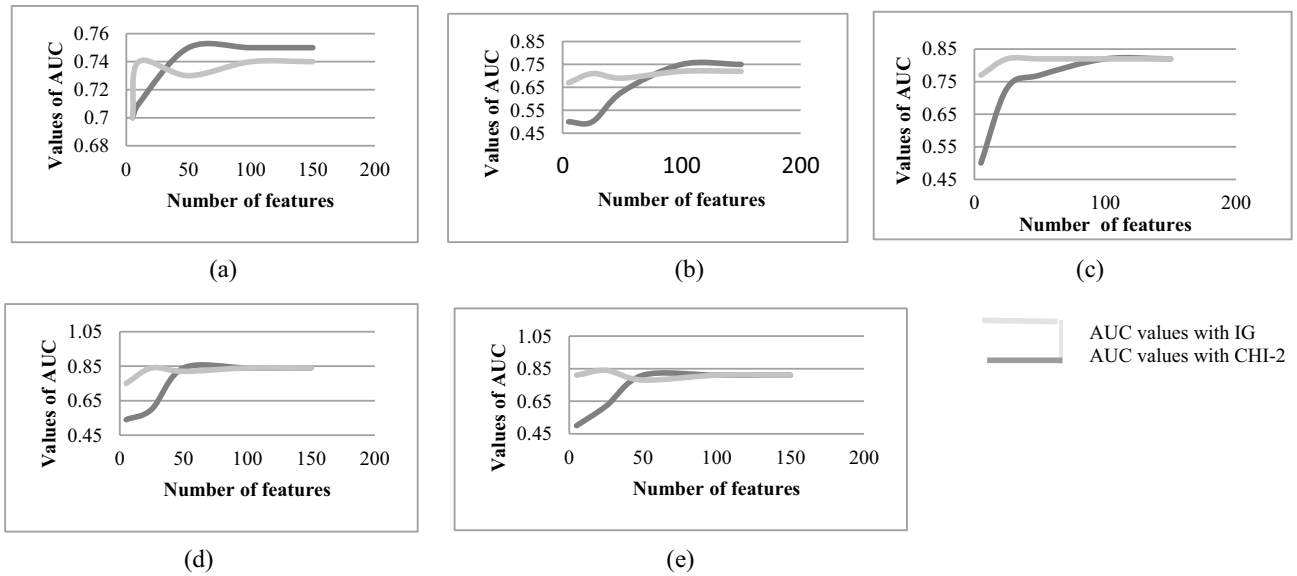


Fig. 5 Relation between AUC values and Number of features in **a** Mozilla Firefox, **b** Mozilla Thunderbird, **c** Mozilla Seamonkey, **d** Mozilla Firefox ESR, **e** Mozilla Thunderbird ESR

Table 11 Result of Friedman test

	Mean rank (low)		Mean rank (medium)		Mean rank (high)		Mean rank (critical)	
	CHI-2	IG	CHI-2	IG	CHI-2	IG	CHI-2	IG
DT	4.63	4.63	4.63	4.63	4.63	4.63	4.63	4.63
SVM	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50
NB	3.75	3.75	3.75	3.75	3.75	3.75	3.75	3.75
Bagging	3.13	3.13	3.13	3.13	3.13	3.13	3.13	3.13
MNB	3.75	3.75	3.75	3.75	3.75	3.75	3.75	3.75
RF	3.75	3.75	3.75	3.75	3.75	3.75	3.75	3.75
Boosting	4.50	4.50	4.50	4.50	4.50	4.50	4.50	4.50

Table 12 Results of the Wilcoxon test

Pair	Medium (CHI-2)	Critical (CHI-2)	Pair	Medium (CHI-2)	Critical (CHI-2)
SVM–DT	0.08	0.786	Bagging–NB	0.138	0.078
NB–DT	**0.043**	**0.042**	MNB–NB	0.414	0.686
Bagging–DT	0.141	0.216	RF–NB	0.078	0.221
MNB–DT	**0.043**	0.144	Boosting–NB	**0.043**	**0.042**
RF–DT	0.08	0.786	MNB–Bagging	0.068	0.138
Boosting–DT	0.285	0.273	RF–Bagging	0.713	0.414
NB–SVM	0.225	0.141	Boosting–Bagging	**0.044**	**0.043**
Bagging–SVM	0.686	0.786	RF–MNB	**0.042**	**0.043**
MNB–SVM	**0.043**	**0.043**	Boosting–MNB	**0.043**	**0.042**
RF–SVM	0.416	0.785	Boosting–RF	**0.044**	**0.041**
Boosting–SVM	0.08	0.068			

The particular pair of ML algorithm where a significant difference in the result values is found according to Wilcoxon signed-rank test are shown in bold

1. There is a possibility that new vulnerabilities are being reported and added to the vulnerability database now but not included in the data collected, which we did initially to carry out the study.
2. For this study, five data sets are used, over which fourteen classification models are being developed. A selection of these can cause biased results obtained in the study.
3. In the literature, there is a diversity of classification models that could have been considered. Still, our selection represents the most widely used classification model, and hence we can claim them to be representing all other models.
4. The results obtained in this study cannot be generalized to the other available datasets.
5. Use of tenfold cross-validation reduced the risk of threat to internal validity.
6. Statistical test reduces the threat to the validity of conclusions by justifying the comparison of all the SVSPM developed properly.
7. For performance evaluation, out of so many parameters, only AUC values are considered, which again may threaten the validity of results.
8. A vulnerability report contains many fields, but our research used only the description portion. The use of limited fields from the available ones can also threaten the validity of the study.

After answering all the RQs, the findings of the result analysis can be summed up as depicted in Table 13.

7 Conclusion and future work

Nowadays, when most human activities depend on different software systems, it becomes crucial to keep in mind their reliability. As most of the activities are carried out online, it poses a threat to exploit various vulnerabilities. Therefore a need arises to automate the task of severity prediction of vulnerabilities and build an SVSPM. SVSPM aims at severity prediction so that appropriate actions can be performed to protect software from exploiting the vulnerability. To achieve this aim, in this study, we focused on vulnerability reports. From these reports, the description is targeted and acted as input to SVSPM. As it contains textual information regarding the vulnerability, text mining concepts are applied to handle this unstructured data. FS is also applied to handle the "curse-of-dimensionality" so that the prediction can be made ahead using various ML algorithms. With the use of two FS techniques and seven ML algorithms, we ended up with the development of fourteen SVSPM. After doing the result analysis, we concluded the following conclusion points:

1. The model performed better for the medium and the critical severity level of the vulnerability. Out of two FS techniques, IG gave better results.
2. Best SVSPM using ML algorithm corresponding to each dataset is found as well. For medium and critical level of severity, SVSPM developed using Boosting algorithm performs the best. Although results were not good for the low and the high level of severity, NB gave comparatively better results.
3. A comparison is also made to identify significant differences among various ML algorithms using Friedman and Wilcoxon Signed Rank test. The results for medium and critical level of severity, when CHI-2 was used for FS technique came out to be significant. Proceeding with pair-wise Wilcoxon signed-rank test, boosting algorithm proved to be the best performing algorithm.
4. It is found that, with the increase of number of features, i.e., N, the performance of SVSPM improves and later on becomes stable or slightly reduced due to the addition of some redundant or irrelevant features.

In the future, there are three possible pathways for further development. Firstly, looking at point 2 of the conclusion, the data which was taken as input is analyzed. Looking at the data, it is found that the data on which we are working is imbalanced data having an unequal number of instances for different levels of severity of vulnerability. Therefore, we aim to solve the issue of imbalanced data, so that better results can be obtained for all levels of severity, including the low and the high level of severity (e.g., Zhou et al. 2018). Different data balancing strategies are used to overcome imbalanced data issues, categorized as data level,

Table 13 Results in a Nutshell

Which FS technique gave better results?	IG
Which level of severity gave better results?	Medium and Critical, followed by Low and High
Which ML algorithm performed better?	At low level, NB algorithm, At medium level, Boosting algorithm At high level, NB algorithm, At critical level, Boosting algorithm
What is the optimum value of N for achieving the best results?	After N = 100, performance becomes stable

algorithmic level, and hybrid strategy for data balancing. The second pathway is to develop a novel technique for FS so that the performance of the SVSPM can be enhanced. As we are working on a textual dataset that generates feature vectors of high dimension, therefore developing a novel FS technique is very crucial to deal with the curse of dimensionality and improve the performance of the prediction model. For this purpose, algorithms such as deep learning algorithms and evolutionary algorithms can be used. These algorithms are gaining popularity in literature and are used in various paradigms (Dong et al. 2018; Alqudah et al. 2021; Wu et al. 2020; Saha et al. 2022; Kaur and Khehra 2022). The third pathway aims toward generalization of the results. For that purpose, other than Mozilla, different datasets can be added, having a high number of vulnerabilities, investigating whether the proposed model gives good performance or not.

Funding No funding was received for this work.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Human and animals rights This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent For this type of study, informed consent is not required.

References

Alqudah AM, Qazan S, Al-Ebbini L, Alquran H, Qasmieh IA (2021) ECG heartbeat arrhythmias classification: a comparison study between different types of spectrum representation and convolutional neural networks architectures. J Ambient Intell Humnaiz Comput 13:1–31

Alves H, Fonseca B, Antunes N (2016) Software metrics and security vulnerabilities: dataset and exploratory study. In: 12th European dependable computing conference (EDCC). Gothenburg, Sweden, pp 37–44. https://doi.org/10.1109/EDCC.2016.34

Babalau I, Corlatescu D, Grigorescu O, Sandescu C, Dascalu M (2021) Severity prediction of software vulnerabilities based on their text description. In: 2021 23rd international symposium on symbolic and numeric algorithms for scientific computing (SYNASC), pp 171–177

Bilge L, Dumitra T (2012) Before we knew it: an empirical study of zero-day attacks in the real world. In: ACM conference on computer and communications security, pp 833–844

Blumberg R, Atre S (2003) The problem with unstructured data. DM Rev 13:42–49

Chen J, Kudjo P, Mensah S, Brown S, Akorfu G (2020) An automatic software vulnerability classification framework using term frequency-inverse gravity moment and feature selection. J Syst Softw 167:110616. https://doi.org/10.1016/j.jss.2020.110616

Coulter R, Han QL, Pan L, Zhang J, Xiang Y (2020) Data-driven cyber security in perspective–intelligent traffic analysis. IEEE Trans Cybern 50(50):3081–3093

Dong F, Wang J, Li Q, Xu G, Zhang S (2018) Defect prediction in android binary executables using deep neural network. Wirel Pers Commun 102:2261–2285

Erturk E, Akcapinar E (2015) A comparison of some soft computing methods for software fault prediction. Expert Syst Appl 42(4):1872–1879. https://doi.org/10.1016/j.eswa.2014.10.025

FIRST (2007) Common vulnerability scoring system (cvss) version 2.0. https://www.first.org/cvss/v2/guide#i1.2. Accessed 4 Oct 2022

Furnell S (2009) Cybercrime in society. Connected minds, emerging cultures: cybercultures in online learning

Grimes S (2008).Unstructured data and the 80 percent rule, Clarabridge Bridgepoints

Gupta MK, Govil MC, Singh G (2015) Text-mining based predictive model to detect XSS vulnerable files in web applications. In: Annual IEEE India conference (INDICON). https://doi.org/10.1109/INDICON.2015.7443332

Han Z, Li X, Xing Z, Liu H, Feng Z (2017) Learning to predict severity of software vulnerability using only vulnerability description. In: IEEE international conference on software maintenance and evolution, pp 125–136. https://doi.org/10.1109/ICSME.2017.52

He H, Garcia EA (2009) Learning from imbalanced data. IEEE Trans Knowl Data Eng 21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239

Huang G, Li Y, Wang Q, Ren J, Cheng Y, Zhao X (2019) Automatic classification method for software vulnerability based on deep neural network. IEEE Access 7:28291–28298. https://doi.org/10.1109/ACCESS.2019.2900462

Kaur J, Khehra BS (2022) Fuzzy logic and hybrid based approaches for the risk of heart disease detection: state-of-the-art review. J Inst Eng (india): Ser B 103(2):681–697

Keefe TM, Koprinska I (2009) Feature selection and weighting methods in sentiment analysis. In: Australasian document computing symposium

Kekul H, Ergen B, Arslan H (2022) A multiclass approach to estimating software vulnerability severity rating with statistical and word embedding methods. Int J Comput Netw Inf Secur 12(4):27

Khazaei A, Ghasemzadeh M, Derhami V (2016) An automatic method for CVSS scores prediction using vulnerabilities description. J Intell Fuzzy Syst 30(1):89–96. https://doi.org/10.3233/IFS-151733

Kudjo PK, Brown SA, Mensah S (2023) Improving software vulnerability classification performance using normalized difference measures. Int J Syst Assur Eng Manag 14:1–18

Kudjo PK, Chen J, Mensah (2020) The effect of Bellwether analysis on software vulnerability severity prediction models. Softw Qual J 28:1413–1446. https://doi.org/10.1007/s11219-019-09490-1

Kumar KH, Srinivas K (2021) Preliminary performance study of a brief review on machine learning techniques for analogy based software effort estimation. J Ambient Intell Humaniz Comput 14:1–25

Le THM, Sabir B, Babar MA (2019) Automated software vulnerability assessment with concept drift. In: IEEE/ACM 16th international conference on mining software repositories (MSR). pp 371–382. https://doi.org/10.1109/MSR.2019.00063

Li X, Ren X, Xue Y, Xing Z, Sun J (2023) Prediction of vulnerability characteristics based on vulnerability description and prompt learning. In: IEEE international conference on software analysis, evolution and reengineering (SANER), pp 604–615

Liu H, Motoda H, Setiono R, Zhao Z (2010) Feature selection: an ever-evolving frontier in data mining. In: Proceedings of the fourth workshop on feature selection in data mining, vol 4, pp 4–13

Liu K, Zhou Y, Wang Q, Zhu X (2019) Vulnerability severity prediction with deep neural network. In: International conference on big

data and information analytics (BigDIA), pp 114–119. https://doi.org/10.1109/BigDIA.2019.8802851
Malhotra R (2016) Empirical research in software engineering: concepts, analysis, and applications, 1st edn. CRC, Boca Raton
McGraw G (2006) Software security: building security, vol 1. Addison-Wesley Professional, Boston
Mehta P, Aggarwal S, Tandon A (2023) the effect of topic modelling on prediction of criticality levels of software vulnerabilities. Informatica 47
Menzies T, Marcus A (2008) Automated severity assessment of software defect reports. In: IEEE International conference on software maintenance, Beijing, pp 346–355. https://doi.org/10.1109/ICSM.2008.4658083
Microsoft C (2002) Microsoft security response center security bulletin severity rating system. https://technet.microsoft.com/zhcn/security/gg309177.aspx. Accessed 4 Oct 2022
Mitre C (2022) Common vulnerabilities and exposures (CVE). https://cve.mitre.org/. Accessed 4 Oct 2022
Ni X, Zheng J, Guo Y, Jin X, Li L (2022) Predicting severity of software vulnerability based on BERT-CNN. In: International conference on computer engineering and artificial intelligence (ICCEAI), pp 711–715
Okutan A, Mirakhorli M (2022) Predicting the severity and exploitability of vulnerability reports using convolutional neural nets. In: Proceedings of the 3rd international workshop on engineering and cybersecurity of critical systems, pp 1–8
Porter M (1980) An algorithm for suffix stripping, program, pp 30–137
Rehman A, Javed K, Babri HA (2017) Feature selection based on a normalized diference measure for text classifcation. Inf Process Manag 53:473–489
Russell R, Kim L, Hamilton L, Lazovich T, Harer J, Ozdemir O, Ellingwood P, McConley M (2018) Automated vulnerability detection in source code using deep representation learning, pp 757–762. https://doi.org/10.1109/ICMLA.2018.00120
Sagi O, Rokach L (2018) Ensemble learning: a survey. Wires. https://doi.org/10.1002/widm.1249
Saha A, Rajak S, Saha J, Chowdhury C (2022) A survey of machine learning and meta-heuristics approaches for sensor-based human activity recognition systems. In: J Ambient Intell Humaniz Comput, pp 1–28
Şahin E, Tosun A (2019) A conceptual replication on predicting the severity of software vulnerabilities. In: EASE '19: proceedings of the evaluation and assessment on software engineering, pp 244–250. https://doi.org/10.1145/3319008.3319033
Sans I (2022) Sans critical vulnerability analysis archive. http://www.sans.org/newsletters/cva/. Accessed 4 Oct 2022
Sharma R, Sibal R, Sabharwal S (2020) Software vulnerability prioritization using vulnerability description. Int J Syst Assur Eng Manag. https://doi.org/10.1007/s13198-020-01021-7
Spanos G, Angelis L (2018) A multi-target approach to estimate software vulnerability characteristics and severity scores. J Syst Softw 146:152–166. https://doi.org/10.1016/j.jss.2018.09.039
Theisen C, Williams L (2020) Better together: comparing vulnerability prediction models. Inf Softw Technol 119:106204
US-CERT (2006) Uscert vulnerability note field descriptions. http://www.kb.cert.org/vuls/html/fieldhelp. Accessed 4 Oct 2022
Vijayarani S, Ilamathi J, Nithya S (2015) Preprocessing techniques for text mining-an overview. Int J Comput Sci Commun Netw 5(1):7–16
Violettas GE, Theodorou TL, Stephanides GC (2013) E-learning software security: tested for security vulnerabilities & issues. E-learning best practices in management, design and development of e-courses. In: Standards of excellence and creativity, pp 233–240
Wang P, Zhou Y, Sun B, Zhang W (2019a) Intelligent prediction of vulnerability severity level based on text mining and XGBboost. In: Eleventh international conference on advanced computational intelligence (ICACI), pp 72–77. https://doi.org/10.1109/ICACI.2019.8778469
Wang P, Zhou Y, Sun B, Zhang W (2019b) Intelligent prediction of vulnerability severity level based on text mining and XGBboost. In: Conference on advanced computational intelligence (ICACI), pp 72–77. https://doi.org/10.1109/ICACI.2019.8778469
Wu JMT, Tsai MH, Xiao SH, Liaw YP (2020) A deep neural network electrocardiogram analysis framework for left ventricular hypertrophy prediction. J Ambient Intell Humaniz Comput 1–17
Yiming Y, Pederson JO (1997) A comparative study on feature selection in text categorization. In: International conference on machine learning, vol 97, pp 412–420
Younan Y (2013) 25 years of vulnerabilities: 1988–2012. Sourcefire vulnerability research team.
Zerkane S, Espes D, Le Parc P, Cuppens F (2017) Vulnerability analysis of software defined networking. In: International symposium on foundations and practice of security, pp 97–116. https://doi.org/10.1007/978-3-319-51966-1
Zhou L, Li R, Zhang S, Wang H (2018) Imbalanced data processing model for software defect prediction. Wirel Pers Commun 102:937–950

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Educational Administration: Theory and Practice
2024, 30(5), 13164-13177
ISSN: 2148-2403
https://kuey.net/

Research Article

The Influence Of Color On Visual Psychology And Cognitive Behavior: A Study In Paediatrics Hospital Environment

Monica Singh[1*], Ranganath M Singari[2], Mihir Bholey[3]

[1*,2]Department of Design, Delhi Technological University, Delhi, India.
[1]https://orcid.org/0000-0003-4537-7648,monicasingh.dtu@gmail.com,[2]https://orcid.org/0000-0003-4981-3857 ranganathdce@gmail.com
[3]National Institute of Design, Gandhi Nagar, India. https://orcid.org/0000-0001-6702-9631 bholey.mihir@gmail.com

***Corresponding Author:** Monica Singh
*Monicasingh.Dtu@Gmail.Com

Citation: Monica Singh et al. (2024), The Influence Of Color On Visual Psychology And Cognitive Behavior: A Study In Paediatrics Hospital Environment, Educational Administration: Theory and Practice, 30(5), 13164-13177
Doi: 10.53555/kuey.v30i5.5679

ARTICLE INFO

ABSTRACT

The comprehensive study on "The Influence of Color on Visual Psychology and Cognitive Behavior: A Study in Paediatrics Hospital Environment" explores the profound influence of color choices on paediatrics hospital management, with a focus on their effects on patients, staff, and families. The research findings reveal that color psychology plays a pivotal role in creating healing environments tailored to the unique needs of children. Warm and vibrant colors, such as yellows and reds, are shown to uplift the mood of young patients in play areas, while softer pastels like blues and greens can soothe and potentially aid in pain perception in recovery rooms. Color psychology extends to staff perceptions, where a well-designed hospital with carefully chosen colors can boost morale, enhance job satisfaction, and foster better communication and teamwork among healthcare professionals. For families, the study highlights the role of colors in alleviating stress and anxiety, advocating for comfortable and relaxing spaces in waiting areas and family rooms.

These findings have critical implications for hospital environment, including potential improvements in patient outcomes, staff satisfaction, and family experiences. The study underscores the need for future research on cultural variations, long-term effects, multi-sensory design, cost-effective solutions, and patient and family involvement in hospital design. Incorporating these insights can lead to more compassionate, supportive, and effective healthcare environments for paediatrics patients and their families.

KEYWORDS: Color Psychology, Paediatrics Hospital, Healing Environments, User Experience, Cognitive Behavior

INTRODUCTION

Paediatrics hospitals are institutions that are crucial for the comprehensive and specialized medical care of infants, children, and adolescents. With a focus on addressing the unique physical, emotional, and developmental needs of young patients, these hospitals play a vital role in promoting the health and well-being of the paediatrics population. From neonates, who are the most vulnerable and delicate, to adolescents on the brink of adulthood, the spectrum of care offered in these hospitals encompasses various stages of growth and development.

In the nurturing environment of a paediatrics hospital, the aim is not just to provide medical treatment but also to create a space that is comfortable, engaging, and conducive to the emotional healing of young patients. Such an environment is carefully crafted, keeping in mind the principles of color psychology and the profound impact that different hues can have on individuals, especially children. Researchers and healthcare

Copyright © 2024 by Author/s and Licensed by Kuey. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

professionals recognize that the atmosphere in which medical treatment is provided can significantly influence the emotional state and well-being of patients, families, and staff. The following objectives were considered for the study.

1. The first objective involves conducting a comprehensive review of the existing literature on color psychology within healthcare environments. This review will place a special emphasis on understanding how colors influence emotions, behavior, and overall well-being. By synthesizing this body of knowledge, we aim to establish a foundation for the subsequent research.
2. The second objective is to closely examine the critical role of color selection and design in the environment of hospitals. This examination will illuminate the connection between color choices and the outcomes experienced by patients. By exploring this relationship, we aim to shed light on the potential benefits of strategic color design in healthcare settings.
3. The third objective recognizes the unique challenges and considerations that arise when designing paediatrics healthcare environments. Creating spaces that are child-friendly and supportive requires a distinct set of guidelines and approaches. Understanding these factors is crucial in ensuring that paediatrics hospitals meet the specific needs of young patients.
4. The fourth objective involves a rigorous investigation into the influence of colors within paediatrics hospital settings. This investigation will rely on a combination of surveys and observational data to comprehensively explore how colors impact the emotions, behaviors, and experiences of patients, healthcare staff, and families. With the accumulated knowledge and insights.
5. The fifth objective is to provide practical color recommendations for the strategic design of paediatrics hospitals. These recommendations will encompass considerations such as age-appropriate color palettes, effective lighting strategies, and overall design principles aimed at optimizing the healing environment.
6. The sixth objective is to illustrate these findings and recommendations through a compelling case study. This case study will spotlight a paediatrics hospital that has implemented color psychology principles in its design and environment.

By assessing the real-world impact of these changes on patients, staff, and families, we aim to provide tangible evidence of the benefits associated with incorporating color psychology into hospital environment.

The significance of this research lies in its potential to contribute to the enhancement of paediatrics healthcare environments, ultimately benefiting the physical and emotional well-being of young patients, their families, and healthcare providers. By exploring the relationship between color and emotions in hospital settings, this study can offer valuable insights for hospital administrators, designers, and healthcare professionals tasked with creating and managing paediatrics healthcare facilities. Furthermore, the findings of this study can have broader implications for the field of hospital management and healthcare design. While the focus is on paediatrics hospitals, the principles of color psychology can be adapted and applied to healthcare settings for adults as well. Understanding how color choices can positively influence patient and staff experiences may lead to improved patient satisfaction, shorter hospital stays, and potentially better treatment outcomes.

LITERATURE REVIEW

The impact of the physical environment in healthcare settings, particularly within the context of children's hospitals, has garnered increasing attention in recent years. The use of color in hospital design has emerged as a pivotal factor affecting patients' well-being, comfort, and overall experiences. This literature review examines various studies that explore the role of color in paediatrics hospital environments, its implications for patient care, and the influence of color psychology on hospital design.

Color Psychology in Hospital Environments: Color psychology, the study of how colors can affect human emotions and behavior, plays a vital role in healthcare settings. In the study conducted by Sara Nourmusavi Nasab, Amir Reza Karimi Azeri, and S. Mirbazel (2020), the authors emphasized that children prefer hospital environments with entertainment facilities. This highlights the significance of creating visually engaging spaces that contribute positively to the emotional state of young patients (Nourmusavi Nasab et al., 2020).

Thematic Color Design for Children's Hospitals: Thematic color design, as discussed by Ali AbuLawi (2021), is deemed appropriate for children of all age ranges. This approach recognizes the importance of thematic and visually engaging environments that can enhance children's experiences in hospitals (AbuLawi, 2021).

Preferences of Thematic Design and Color: The preference of thematic designs and colors among children and young people was explored by J. Coad and N. Coad (2008). Their study, published in the Journal of Child Health Care, indicated that children and young people in a new children's unit preferred thematic designs of walls, doors, and floors. This underscores the importance of tailoring the hospital environment to meet the preferences of its younger occupants (Coad & Coad, 2008).

Psychological Wellbeing of Paediatrics Patients: In the research conducted by Valentina Marques da Rosa, P. Brust-Renck, and L. Tonetto (2022), the authors highlighted the impact of the hospital environment on all components of children's psychological well-being. This underscores the potential for hospital design, including color choices, to support the emotional needs of paediatrics patients (Marques da Rosa et al., 2022).

Color Environment Design in Paediatrics Care Units: Color selection in paediatrics hospital design was explored by L. Hong-xia (2011). The study suggested that colors are associated with specific emotions in paediatrics hospitals, emphasizing the need for thoughtful color choices in healthcare environments (Hong-xia, 2011).

Color Perception in Paediatrics Patient Rooms: A study by Jin Gyu "Phillip" Park (2009) compared the color preferences of healthy children and paediatrics outpatients. Healthy children exhibited higher mean scores of color preferences, underscoring the importance of considering color psychology in the design of paediatrics patient rooms (Park, 2009).

Appropriate Colors in Children's Rooms: Ajilian Maryam and T. Ali (2014) emphasized the diverse impacts of color, including its ability to inspire, soothe, heal, and agitate. Their study highlighted the multifaceted role of color in paediatrics hospital environments, suggesting its potential to create varied emotional responses (Maryam & Ali, 2014).

Child-Friendly Hospital Architecture: A child-friendly hospital environment was explored by Laure Verschoren, M. Annemans, I. V. Steenwinkel, and A. Heylighen (2015). Their research highlighted the importance of a child-friendly atmosphere that supports daily life, both socially and spatially, reinforcing the notion that hospitals should be designed with the unique needs of children in mind (Verschoren et al., 2015).

Role of Hospital Design in Reducing Anxiety: J. Cartland, H. Ruch-Ross, L. Carr, Audrey E. Hall, R. Olsen, Ellen Rosendale, and Susan Ruohonen (2018) investigated the impact of hospital design on anxiety levels among paediatrics patients. Children in the new facility experienced less anxiety than in the old facility, emphasizing the therapeutic potential of well-designed hospital spaces (Cartland et al., 2018).

Interactive Design in Children's Hospitals: The study conducted in 2022 emphasized the importance of interactive design in children's hospitals, suggesting that children feel much better in environments that incorporate emotional factors through interactive design (Anonymous, 2022).

Cross-Cultural Perspectives on Color Preferences: Jin Gyu "Phillip" Park and Changbae Park (2013) explored the differences in color preferences between American and Korean paediatrics patients. Korean paediatrics patients showed significantly higher preference scores for white than their American counterparts, underscoring the influence of culture on color perceptions in healthcare environments (Park & Park, 2013).

Light and Color in Healthcare Environments: Research by Navid Khaleghimoghaddama (2023) emphasized the importance of the proper use of color and light in healthcare environments. These elements were found to have a significant impact on healthcare outcomes, suggesting their potential to enhance the healing process (Khaleghimoghaddama, 2023).

Comprehensive Impact of Color in Healthcare: H. Ghamari and Cherif Amor (2016) summarized the extensive effects of color in healthcare environments. Their research highlighted color's role in reducing medical errors, promoting well-being, reducing stress, improving sleep, shortening hospital stays, enhancing spatial orientation, increasing patient satisfaction, and improving staff morale and productivity (Ghamari & Amor, 2016).

Color Selection and Patient Well-Being: Arezou Eshaghabadi, Pirhossein Koulivand, and H. Kazemi (2017) explored the impact of color selection on patient well-being in hospital design. Their study emphasized the significance of color choices in contributing to patients' overall experiences and outcomes (Eshaghabadi et al., 2017).

Environmental Graphics in Children's Wards: The analysis conducted by L. Jeddi, Fereidoon Kasrayee, S. Jeddi, and M. Taghipour (2016) focused on the role of suitable color in graphical design in children's hospital wards. Their research highlighted the importance of coordination between architectural and graphical design to improve the hospital's atmosphere and quality of care for children (Jeddi et al., 2016).

Environmental Psychology in Children's Hospitals: Lu Feng-hu (2010) discussed the association of colors with specific emotions in the decorative design of children's hospitals, underscoring the potential psychological impact of color choices in healthcare settings (Feng-hu, 2010).

Cultural Significance of Color: Jain Kwon (2010) explored the cultural meaning of color in healthcare environments, emphasizing the importance of considering cultural perspectives when choosing colors for hospital design (Kwon, 2010).

Lighting and Color in Hospital Design: Research by Hilary Dalke, Jenny Little, Elga Niemann, Nilgun Camgoz, Guillaume Steadman, Sarah Hill, and Laura Stott (2006) emphasized the importance of a well-balanced and attractive environment in patient health. Their study highlighted the potential for color and lighting to positively affect patient recovery and staff performance (Dalke et al., 2006).

Quality of the Visual Hospital Environment: H. Dalke, P. Littlefair, and D. Loe (2004) discussed how the quality of the visual hospital environment can have a positive psychological effect on patient recovery and staff performance, reinforcing the importance of visual elements like color in healthcare design (Dalke et al., 2004).

Children's Emotional Associations with Colors: A study by C. Boyatzis and Reenu Varghese (1994) explored children's emotional associations with colors. Their findings revealed that children's emotional reactions to bright colors became increasingly positive with age, suggesting the developmental nuances in color preferences among paediatrics patients (Boyatzis & Varghese, 1994).

Colors in Public Places: Parisa Zraati (2013) discussed the influence of colors in public places, emphasizing that colors can affect users. This highlights the significance of color choices in spaces like hospital waiting areas (Zraati, 2013).

This literature review highlights the multifaceted role of color in paediatrics hospital environments, emphasizing its potential to influence the emotional state, preferences, and well-being of young patients. The studies reviewed collectively underscore the importance of thoughtful color selection and design in creating hospital spaces that are conducive to healing, comfort, and the overall well-being of paediatrics patients and their families.

The Significance of Color Psychology in Paediatrics Hospital Design

The impact of color psychology in paediatrics hospital management, certain studies suggest that colors can have psychological effects on both children and adults. In the context of paediatrics hospitals, the use of certain colors can potentially contribute to creating a more comforting and soothing environment for young patients. For example, some hospitals use calming, warm colors such as light blues, greens, and pastels to create a relaxed atmosphere, which can help reduce stress and anxiety in children. These colors are often associated with feelings of tranquility and can have a positive impact on the emotional well-being of young patients. Additionally, some paediatrics hospitals might use vibrant and cheerful colors in play areas or waiting rooms to create an engaging and lively environment for children. This approach aims to promote a sense of joy and positivity, potentially aiding in the overall healing process and creating a more child-friendly atmosphere.

Fig 1. Centre for Children - Kokilaben Dhirubhai Ambani Hospital
https://www.kokilabenhospital.com/departments/centresofexcellence/centrefor_children.html

Fig 2. Manipal Paediatric Hospital Bangalore
https://www.manipalhospitals.com/specialities/paediatric-surgery

Fig 3. Bhagirathi Neotia Woman & Child Care Centre in Kolkata
https://www.haarway.com/bhagirathi-neotia-woman-child-care-centre-near-opposite-assembly-of-god-church-school-kolkata-contact-phone-number-address

METHODOLOGY

This study employs a mixed-methods research design to analyze the impact of color on pediatric healthcare environments. By integrating quantitative and qualitative approaches, the study aims to comprehensively understand color preferences, emotional associations, and perceived environmental effects from the perspectives of hospital staff, young patients, and their families/visitors.

A. The research seeks to answer the following questions:
1. How do different colors in hospital environments affect the emotional well-being of pediatric patients?
2. What specific colors are perceived as most comforting by children in a hospital setting?
3. How does healthcare staff perceive the impact of color on their work environment and stress levels?
4. What are the experiences of families regarding the color schemes in pediatric hospitals?
5. How does the use of color in pediatric hospital design influence the behavior of young patients during their stay?
6. What are the cultural considerations in choosing colors for pediatric healthcare environments?
7. How do adolescents perceive color differently from younger children in hospital settings?
8. What role does natural lighting play in enhancing the effects of color in pediatric hospitals?
9. How do different age groups respond to various color palettes in a healthcare setting?
10. What are the key design elements, aside from color, that contribute to a healing environment for children in hospitals?
11. How do pediatric patients' emotional responses to color compare with those of their parents or guardians?
12. What are the most significant challenges faced by designers when implementing color psychology principles in pediatric hospitals?
13. How do colors impact the perceived safety and security within pediatric hospital spaces?
14. What feedback do pediatric patients and their families provide about color schemes during their hospital experience?
15. How do specific colors in pediatric hospital rooms influence the length of stay or recovery time for young patients?
16. What are the psychological effects of color changes over different times of the day in a pediatric hospital setting?
17. How do colors in waiting areas versus treatment rooms affect patient anxiety levels?
18. What are the perceived benefits of incorporating nature-inspired color schemes in pediatric hospitals?
19. How do different colors in hospital environments affect the interaction and communication between patients and healthcare staff?
20. What case studies or examples illustrate the successful implementation of color psychology in pediatric hospital design, and what lessons can be learned from them?

B. Data will be collected from three groups: hospital staff, young patients, and families/visitors using surveys, interviews and observations.

Surveys

1. Participants: Hospital staff, young patients, and families/visitors.
2. Instruments: Structured questionnaires developed based on the research questions.
3. Variables: Color preference, emotional association, and perceived effect on the environment.

Interviews

1. Participants: Selected individuals from each group to gain in-depth insights.
2. Instruments: Semi-structured interview guides.

Observations

1. Setting: Pediatric hospital environments.
2. Focus: Interaction with color schemes in different areas (e.g., play areas, recovery rooms, waiting areas).

C. Data analysis will be done using the available data.

The study aims to provide evidence-based recommendations for the use of color in pediatric hospital environments to enhance emotional well-being, improve staff morale, and support families. Insights from this study will inform hospital design strategies, contributing to more effective and compassionate pediatric healthcare settings. This methodology ensures a comprehensive and systematic approach to understanding the complex impacts of color on the pediatric healthcare environment, facilitating data-driven design decisions.

RESULT AND ANALYSIS

From the moment a child steps into the hospital, they are greeted with a lively and colorful setting. Playful and vibrant toys adorn the waiting areas, creating a sense of wonder and distraction. The walls are painted in soothing hues, carefully selected to instill a feeling of calm and reassurance, mitigating the anxiety often associated with medical settings. These deliberate color choices are not random but are based on a deep understanding of the psychological impact of colors on human emotions.

Moreover, the importance of color psychology extends beyond the visual appeal. Studies have shown that specific color schemes can significantly influence mood, behavior, and even physical well-being. In paediatrics hospitals, this understanding translates into the strategic use of colors to facilitate healing and improve the overall hospital experience for young patients. In recovery rooms, where children recuperate from procedures and surgeries, the emphasis is on creating a serene and tranquil environment. Soft, pastel shades of blue and green are often used, as these colors are known to evoke feelings of serenity and relaxation. This calming effect can help reduce stress and discomfort, contributing to a more positive recovery experience. By integrating these subtle yet impactful color schemes, paediatrics hospitals aim to create a supportive and nurturing space where children can heal both physically and emotionally.

Fig 4. Maternal & Child Health Care Room

Fig 5. Imaging Room for Childrens

Fig 6. Paediatric Treatment Room

Fig 7. Paediatric Intensive Care Unit (PICU)

Contrastingly, in play areas and activity zones, vibrant and energetic colors take center stage. Bright reds, yellows, and oranges stimulate energy and enthusiasm, fostering a lively and joyful ambiance. These vibrant settings are designed to encourage engagement, promote social interaction, and uplift the spirits of young patients. By immersing children in an environment filled with stimulating colors, paediatrics hospitals aim to create a positive and uplifting atmosphere that aids in the emotional well-being and overall recovery of the patients.

Furthermore, the impact of color psychology extends to the well-being of the hospital staff and families of the patients. By creating an environment that is visually appealing and emotionally supportive, paediatrics hospitals foster a sense of positivity and optimism among the caregivers, contributing to a more conducive work atmosphere.

It should be understood from the above that the integration of color psychology in paediatrics hospitals serves as a powerful tool in creating a holistic healing environment for young patients. Through careful consideration and implementation of color schemes, these hospitals strive to enhance the well-being and comfort of children, their families, and the healthcare professionals involved in their care. By leveraging the psychological impact of colors, paediatrics hospitals continue to pave the way for a more compassionate and effective approach to paediatrics healthcare.

Moreover, the study highlights that the impact of color extends to the hospital staff. A well-designed hospital with carefully chosen colors can foster a positive work atmosphere, leading to higher staff morale and job satisfaction. Colors can also be used for wayfinding, making it easier for staff to navigate the hospital efficiently and collaborate effectively. Families, too, benefit from the thoughtful application of color psychology in paediatrics hospitals. Comfortable and relaxing spaces with soft, neutral colors in waiting areas and family rooms can significantly contribute to their well-being. Muted blues and greens, in particular, help alleviate the stress and anxiety that often accompany a child's hospitalization.

Paediatrics hospitals are indeed playful and colorful, but their vibrant design elements are not just for show. They are a manifestation of the intricate relationship between color psychology and hospital Environment. The careful selection of colors contributes to a healing environment that supports the emotional needs of young patients, ensures staff satisfaction, and eases the burden on families during challenging times.

Healthcare environments, by their very nature, are spaces of vulnerability and emotional sensitivity, where patients and their families grapple with illness, uncertainty, and distress. As the healthcare industry continues to evolve, one critical aspect gaining recognition for its influence on patient well-being and outcomes is the physical environment itself. This environment encompasses various elements, including architecture, design, and spatial organization, but it is the use of color that forms a subtle yet powerful aspect of the hospital's atmosphere. In particular, paediatrics hospitals, dedicated to caring for children and adolescents, face the unique challenge of creating spaces that are not only medically efficient but also emotionally supportive, engaging, and calming for their young patients and their families. This study delves into the multifaceted role of color psychology within the context of hospital management, with a particular emphasis on its application in paediatrics healthcare settings.

The physical environment in healthcare settings has long been recognized for its potential to impact patient outcomes. Florence Nightingale, the founder of modern nursing, highlighted the importance of the environment in her groundbreaking work, "Notes on Nursing," published in 1859. She emphasized the significance of clean, well-ventilated spaces and suggested that the surroundings themselves could contribute to or hinder the process of healing. Over the decades, this understanding has evolved, leading to a growing awareness of how various environmental factors, including lighting, layout, noise, and aesthetics, influence the emotional state and overall well-being of patients and healthcare providers. In recent years, the healthcare industry has placed increasing importance on the concept of patient-centered care. This approach recognizes the patient as an active participant in their healthcare journey, and it seeks to provide an environment that not only promotes physical healing but also addresses emotional and psychological needs. In this context, the role of color psychology has gained prominence as an essential element in designing healthcare spaces that support the holistic well-being of patients.

Table 1: Color Palette in Paediatrics Hospital Design

Name	Hex Code	Color	Usage
Sky Blue	#87CEEB		Recovery Room Walls
Soft Green	#98FB98		Recovery Room Accents
Sunshine	#FFD700		Playroom Walls
Bubblegum	#FF69B4		Playroom Accents
Serene Gray	#D3D3D3		Waiting Areas and Corridors
Sunny Yellow	#FFFF00		Paediatrics Room Doors

As this study suggests, the power of color in paediatrics healthcare environments is profound, enriching the lives of those it touches.

The purpose of this study is to investigate and elucidate the role of color psychology in hospital management, with a particular focus on its application in paediatrics hospitals. It aims to explore the various ways in which color can influence the emotional state, comfort, and overall experience of young patients, their families, and healthcare staff. By delving into the science of color psychology and its practical application, this study seeks to provide hospital administrators, designers, and healthcare professionals with evidence-based insights that can inform the strategic use of color in paediatrics healthcare environments. This study is driven by several specific objectives that collectively guide the investigation into the relationship between color psychology and hospital management, particularly in the context of paediatrics healthcare settings.

The table 2 provides a reference for the association between emotions and colors. Common colors associated with happiness include bright yellow and red, while calmness is often linked to light blue and green. Vibrant red and orange are associated with excitement, while soft green and blue represent tranquility. Anxiety is often linked to bright red and yellow, while serenity is associated with pale blue and lavender. Bright orange and pink are associated with energy, while warm yellow and beige represent comfort.

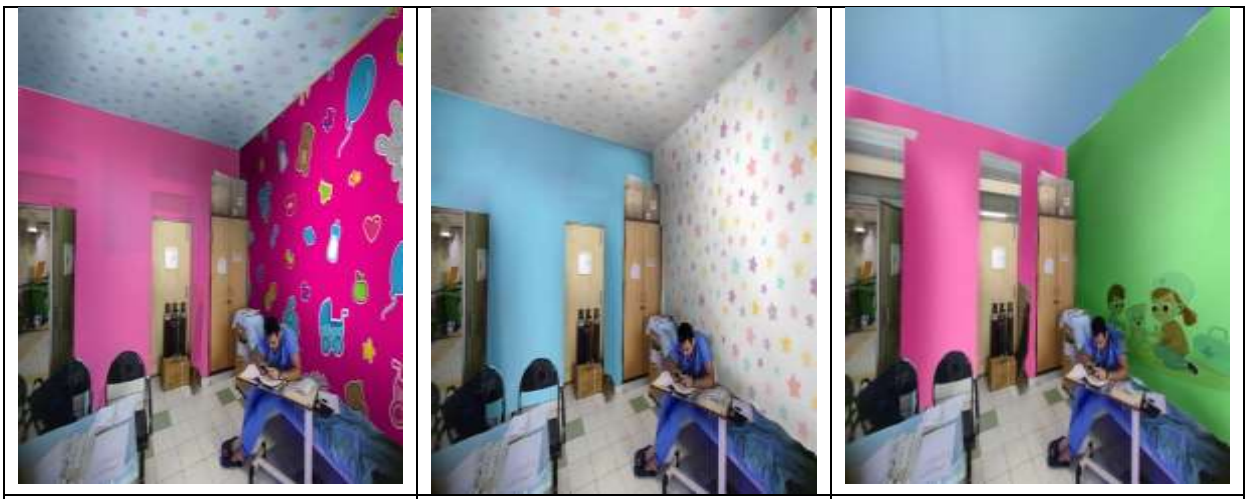
Table 2: Illustrating emotion-color associations based on color psychology

Emotion	Associated Color	References
Happiness	Bright Yellow, Red	Stumpf, C. (2014). Color psychology: Everything you need to know. AVA Publishing.
Calmness	Light Blue, Green	Russell, G. W., & Brooks, D. L. (2004). Color and emotions: A review and meta-analysis. Journal of Experimental Psychology: General, 133(1), 154.
Excitement	Vibrant Red, Orange	Ekman, P., & Davidson, I. (1991). The nature of emotions: Fundamental questions. Oxford University Press.
Tranquillity	Soft Green, Blue	Harrington, T. L., & Peckham, S. L. (2010). The effects of color on emotions: A review. Color Research & Application, 35(4), 211-222.
Anxiety	Bright Red, Yellow	Eysenck, M. W., & Keane, T. M. (2015). Cognitive psychology: A student's handbook (7th ed.). Psychology Press.
Serenity	Pale Blue, Lavender	Moss, R. (2005). The art of tranquility: Meditation and breathing techniques to calm your mind and body. New World Library.
Energy	Bright Orange, Pink	Aticke, D., & Russell, G. W. (2003). The effects of color on perceived energy and affect. Color Research & Application, 28(4), 247-252.
Comfort	Warm Yellow, Beige	Fairchild, M. D. (2005). Munsell color science. Wiley.

The subsequent visuals depict envisioned color schemes implemented for particular healthcare rooms, inspired by the significant insights derived from the study titled " The Influence of Color on Visual Psychology and Cognitive Behavior: A Study in Paediatrics Hospital Environment ." These proposals are crafted to leverage the constructive influence of color psychology, with the goal of elevating the overall well-being and experiences of patients, families, and healthcare professionals within specific medical environments.

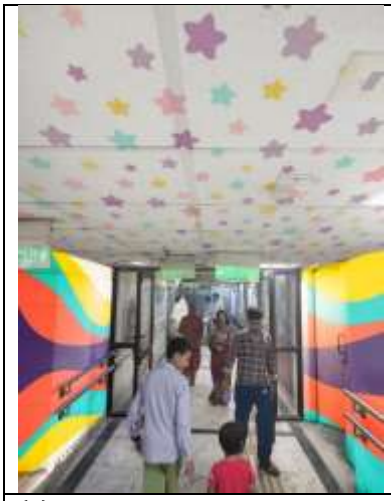
In alignment with the nurturing essence of maternal and child health care, the recommended color palette embraces soft pastels such as calming blues and gentle greens. Introducing accents of warm tones, such as light yellows or peach, is suggested to establish a comforting and supportive ambiance, fostering a sense of care and reassurance. To transform the imaging room into a more child-friendly space, an energetic and playful color scheme is proposed. The incorporation of bright, cheerful colors like blues, greens, and yellows is envisioned to contribute to a visually engaging atmosphere. Additionally, the introduction of child-friendly themes, such as murals or artwork, is suggested to enhance the overall welcoming nature of the room.

For paediatric treatment rooms, the suggested color scheme amalgamates soothing and uplifting tones. Soft blues and greens are recommended to promote a calming effect, while bursts of warmer tones, such as light oranges or soft purples, aim to maintain a positive and supportive ambiance. This approach is designed to create a therapeutic environment tailored to the unique needs of paediatric patients.



(a) (b) (c)

Fig 8 Redesigned Paediatric Treatment Room


(a)

(b)

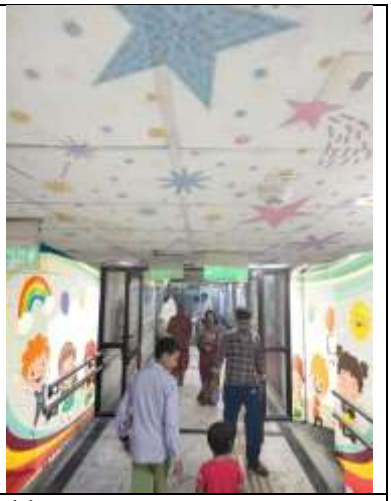

(c)

Fig 9 Redesigned Paediatric room gallery

These proposed color schemes take into account the psychological impact of colors on emotional well-being, with the intention of contributing to more compassionate, supportive, and effective healthcare environments for both patients and their families. It is crucial to evaluate these suggestions within the context of the hospital's overall design scheme, considering practicality and cultural preferences. Input from stakeholders is also essential to ensure a holistic and patient-centric approach to healthcare space design.

(a)

(b)

Fig 10 Redesign Maternal & Child Health Care Room

The results for the analysis of color preference, emotional association, and perceived effect on the environment for the three groups: Hospital Staff, Young Patients, and Families/Visitors.

Table 3: Color Preference Analysis

For Hospital Staff:	For Young Patients:	For Families/Visitors:
Mean Color Preference: 3.0 Median Color Preference: 3 Mode Color Preference: 2 (Green)	Mean Color Preference: 2.3 Median Color Preference: 2.5 Mode Color Preference: 1 (Blue)	Mean Color Preference: 2.58 Median Color Preference: 2.5 Mode Color Preference: 1 (Blue)

Table 4: Emotional Association Analysis

For Hospital Staff:	For Young Patients:	For Families/Visitors:
Mean Emotional Association: 2.48 Median Emotional Association: 2.5 Mode Emotional Association: 2 (Happy)	Mean Emotional Association: 2.52 Median Emotional Association: 3 Mode Emotional Association: 2 (Happy)	Mean Emotional Association: 2.32 Median Emotional Association: 2.5 Mode Emotional Association: 2 (Happy)

Table 5: Perceived Effect on Environment Analysis

For Hospital Staff:	For Young Patients:	For Families/Visitors:
Mean Perceived Effect: 1.38 Median Perceived Effect: 2 Mode Perceived Effect: 2 (Neutral)	Mean Perceived Effect: 1.42 Median Perceived Effect: 2 Mode Perceived Effect: 2 (Neutral)	Mean Perceived Effect: 1.44 Median Perceived Effect: 2 Mode Perceived Effect: 2 (Neutral)

The comprehensive study on the impact of color psychology in paediatrics hospital environment reveals significant insights into how color choices influence the experiences of patients, staff, and families in healthcare environments. Hospital Staff showed a relatively higher mean color preference score of 3.0, indicating a preference for the chosen color palette. Young Patients had a slightly lower mean color preference score of 2.3, suggesting a preference for different colors compared to the staff. Families/Visitors had a mean color preference score of 2.58, aligning more closely with Young Patients. The most preferred color across all groups

was Blue, indicating its universal appeal. Hospital Staff, Young Patients, and Families/Visitors predominantly associated the selected colors with positive emotions, with 'Happy' being the most common emotional association. This positive emotional association suggests that the chosen colors have a generally uplifting effect on the mood of all groups.

Table 6: summarizing the findings of the study, highlighting the impact of colors on young patients, hospital staff, and families in paediatrics hospitals

Aspect	Impact on Young Patients	Impact on Hospital Staff	Impact on Families/Visitors
Emotional Well-being	Engaging and comforting atmosphere, alleviating anxiety during medical procedures and hospital stays.	Positive work atmosphere, potentially higher staff morale and job satisfaction.	Comfortable and relaxing spaces, stress and anxiety reduction during a child's hospitalization.
Atmosphere	Enhances overall atmosphere, promotes positive emotions and boosts mood.	Fosters a positive work environment, aiding staff morale and satisfaction.	Creates a supportive and inviting environment for families and visitors.
Wayfinding	Improves wayfinding for young patients, making navigation more intuitive.	Enhances wayfinding for staff, leading to efficient hospital operations.	Eases navigation and reduces stress for families trying to find their way within the hospital.
Color Preferences	Preferences for bright and cheerful colors in play areas.	Preferences for a well-thought-out color palette in the workplace.	Preferences for soft, neutral colors in waiting areas and family rooms.
Overall Impact	Creates an engaging and comforting environment for paediatrics patients.	Fosters a positive work atmosphere and supports staff well-being.	Provides a welcoming and stress-reducing atmosphere for families and visitors.

CONCLUSION

The findings have several implications for hospital environment, design, and the overall well-being of individuals involved.

The study highlights the potential for color psychology to enhance the experience of young patients in paediatrics hospitals. Colors like Blue, which are commonly preferred, can be strategically used to create more engaging and comforting environments in play areas. Well-designed hospitals with carefully chosen colors can foster a positive work atmosphere for healthcare professionals. This can lead to higher staff morale, job satisfaction, better communication, and effective teamwork. Comfortable and relaxing spaces with appropriate color choices in waiting areas and family rooms can significantly alleviate stress and anxiety for families during a child's hospitalization. The finding that Blue is a universally preferred color suggests its versatility in creating positive environments in paediatrics hospitals. The generally neutral perceptions of the environment by all groups indicate that the chosen colors do not have strong negative effects, which can be considered a positive outcome.

Paediatrics hospitals should consider a strategic approach to color design, tailoring it to specific areas like play zones, recovery rooms, and waiting areas to meet the emotional needs of patients, staff, and families. Collaboration between healthcare professionals, designers, and color psychologists is crucial in selecting the most suitable color palettes for paediatrics hospital environments.

Color Recommendations for Paediatrics Hospitals

Color Recommendations for Paediatrics Hospitals involves selecting and implementing color schemes that support the emotional well-being of young patients, enhance the work environment for hospital staff, and create welcoming spaces for families and visitors. Based on the study's findings:

1. Emotional Well-being: Consider using vibrant and engaging colors in play areas to alleviate anxiety during medical procedures and hospital stays. Soft, soothing colors in recovery rooms can contribute to a calming environment for young patients, aiding pain perception and recovery.
2. Atmosphere: Create a positive and mood-boosting atmosphere in paediatrics hospitals with strategically chosen colors. Incorporate bright and cheerful colors in spaces dedicated to play and recreation. Opt for a well-thought-out color palette that fosters a positive work environment for hospital staff. In waiting areas and family rooms, use soft, neutral colors to provide comfort and reduce stress for families and visitors.
3. Wayfinding: Utilize color as a tool for wayfinding within the hospital. Differentiate departments or areas with distinct colors to make navigation more intuitive for young patients. Similarly, use color for wayfinding cues for hospital staff to enhance operational efficiency. For families and visitors, color can play a role in easing navigation and reducing stress.
4. Color Preferences: Take into account the preferences of various groups within the hospital. Young patients often prefer bright and cheerful colors, while hospital staff may appreciate a well-designed color palette that contributes to a positive work atmosphere. Families and visitors tend to prefer soft and neutral colors in spaces where they seek comfort and support.

5. Play Areas: Utilize bright and vibrant colors like yellows, reds, and playful pastels to create engaging and cheerful play areas. These colors can help uplift the mood of young patients and promote a sense of excitement and wonder.
6. Patient Rooms: Opt for softer pastels, such as calming blues and soothing greens, in patient rooms. These colors can create a peaceful and serene ambiance, potentially aiding in pain perception and recovery. Pastels are also versatile and can be used as a base for other design elements like murals or decals.
7. Hallways and Common Areas: Maintain a cohesive and welcoming atmosphere in hallways and common areas with a harmonious color palette that includes both vibrant and calming colors. This balance ensures that the hospital environment remains engaging but not overly stimulating.
8. Nature-Inspired Palette: Consider incorporating nature-inspired colors, such as soft sky blues, leafy greens, and earthy tones, into various hospital spaces. These colors can help connect young patients to the natural world, providing a sense of comfort and tranquility.
9. Proper Lighting: Enhance the impact of colors by ensuring proper lighting in different areas. Soft, diffused lighting can complement calming colors, creating a soothing atmosphere. In play areas, consider dynamic lighting options that can adapt to different activities and moods.
10. Texture and Depth: Use texture to add depth and richness to the color scheme. Textured wallpapers, upholstery, or decorative elements can enhance the sensory experience and make the environment more engaging for children.
11. Balancing Stimulation and Calmness: Maintain a thoughtful balance between stimulating and calming spaces. While play areas and activity zones can feature brighter and more energetic colors, quieter spaces like patient rooms and family waiting areas should prioritize calming hues. This balance ensures that the hospital environment caters to the diverse emotional needs of young patients.
12. Customization: Consider allowing young patients to personalize their immediate surroundings with removable, washable, or easily changeable elements like posters or decals. This customization can give children a sense of control and comfort during their hospital stay.
13. Multisensory Design: Explore multisensory design elements, such as interactive walls with color-changing LEDs or tactile surfaces, to engage children and provide a sensory-rich experience that aids in distraction and comfort.

These color recommendations for paediatrics hospitals aim to create a healing environment that supports the emotional well-being of young patients, enhances staff morale, and eases the stress experienced by families.
By carefully selecting and implementing colors in different areas of the hospital, healthcare facilities can create a more compassionate, supportive, and effective healthcare environment for paediatrics patients and their families.

REFERENCES

1. Allahyar, M., & Kazemi, F. (2021). Landscape preferences of a children's hospital by children and therapists. Urban Forestry & Urban Greening, 64, 126984. DOI: 10.1016/J.UFUG.2021.126984
2. Alzoubi, H. H., & Al-Rqaibat, S. M. (2015). The effect of hospital design on indoor daylight quality in children section in King Abdullah University Hospital, Jordan. Solar Energy, 117, 83-94. DOI: 10.1016/J.SCS.2014.08.008
3. Aticke, D., & Russell, G. W. (2003). The effects of color on perceived energy and affect. Color Research & Application, 28(4), 247-252.
4. Bubna, K., Hegde, S., & Rao, D. (2017). Role of Colors in Paediatrics Dental Practices. Journal of Clinical Paediatrics Dentistry, 41(3), 193-197. DOI: 10.17796/1053-4628-41.3.193
5. Cartland, J., Ruch-Ross, H., Carr, L., Hall, A. E., Olsen, R., Rosendale, E., & Ruohonen, S. (2018). The Role of Hospital Design in Reducing Anxiety for Paediatrics Patients. HERD, 11(3), 19-30. DOI: 10.1177/1937586718779219
6. Dalke, H., Little, J., Niemann, E., Camgoz, N., Steadman, G., Hill, S., & Stott, L. (2006). Colour and lighting in hospital design. Optics & Laser Technology, 38(6), 343-365. DOI: 10.1016/J.OPTLASTEC.2005.06.040
7. Dalke, H., Littlefair, P., & Loe, D. (2004). Lighting And Colour For Hospital Design. Lighting Research & Technology, 36(4), 349-352.
8. Eisen, S., Ulrich, R., Shepley, M., Varni, J., & Sherman, S. (2008). The stress-reducing effects of art in paediatrics health care: art preferences of healthy children and hospitalized children. Journal of Child Health Care, 12(3), 173-190. DOI: 10.1177/1367493508092507
9. Ekman, P., & Davidson, I. (1991). The nature of emotions: Fundamental questions. Oxford University Press.
10. Eshaghabadi, A., Koulivand, P., & Kazemi, H. (2017). Color Psychology and its Effect on Hospital Design and Patient Treatment. Journal of Educational Technology Research, 2017(2). DOI: 10.18869/ACADPUB.SHEFA.5.2.130
11. Eysenck, M. W., & Keane, T. M. (2015). Cognitive psychology: A student's handbook (7th ed.). Psychology Press.
12. Fairchild, M. D. (2005). Munsell color science. Wiley.

13. Ghamari, H., & Amor, C. (2016). The Role of Color in Healthcare Environments, Emergent Bodies of Evidence-based Design Approach. Science Arena, 4(11), 1-11. DOI: 10.13189/SA.2016.041109
14. Harrington, T. L., & Peckham, S. L. (2010). The effects of color on emotions: A review. Color Research & Application, 35(4), 211-222.
15. Hong-xia, L. (2011). Research on Color Environment Design of Hospital's Paediatrics Care Unit. Procedia Engineering, 11, 344-350.
16. Kotzer, A., Zacharakis, S. K., Raynolds, M., & Buenning, F. (2011). Evaluation of the Built Environment: Staff and Family Satisfaction Pre- and Post-Occupancy of the Children's Hospital. HERD, 4(4), 109-127. DOI: 10.1177/193758671100400405
17. Kwon, J. (2010). Cultural meaning of color in healthcare environments: A symbolic interaction approach. HERD, 3(2), 27-36.
18. Marques da Rosa, V., Brust-Renck, P., & Tonetto, L. (2022). Designing Hospital Environments to Improve the Psychological Wellbeing of Paediatrics Patients. Children, Youth and Environments, 31(3), 98-117. DOI: 10.7721/chilyoutenvi.31.3.0098
19. Moss, R. (2005). The art of tranquility: Meditation and breathing techniques to calm your mind and body. New World Library.
20. Naja, M., Brée, J., & Zaichkowsky, J. (2011). The Use of Ambient Scent to Improve Children's Hospital Experience. In: Camgoz, N., Steadman, G., Hill, S., & Stott, L. (Eds.), Ambient Intelligence Perspectives, 49-57.
21. Park, J. G. "Phillip" (2009). Color Perception in Paediatrics Patient Room Design: Healthy Children vs. Paediatrics Patients. HERD, 2(3), 42-55. DOI: 10.1177/193758670900200302
22. Pearson, M., Gaines, K., Pati, D., Colwell, M., Motheral, L., & Adams, N. (2018). The Physiological Impact of Window Murals on Paediatrics Patients. HERD, 11(3), 65-75. DOI: 10.1177/1937586718800483
23. Russell, G. W., & Brooks, D. L. (2004). Color and emotions: A review and meta-analysis. Journal of Experimental Psychology: General, 133(1), 154.
24. Schuschke, G., & Christiansen, H. (1994). Patient-related color preference and color design in the hospital. Zentralblatt fur Hygiene und Umweltmedizin = International journal of hygiene and environmental medicine, 196(4-5).
25. SilavUtkan, M. (2012). Children Hospital Design in Children Picture. Procedia - Social and Behavioral Sciences, 51, 930-934. DOI: 10.1016/J.SBSPRO.2012.08.127
26. Singh, M., & Singari, R. M. (2023 Color As Expression: Exploring Fauvism's Revolutionary Use Of Vibrant Hues And Its Enduring Impact On Modern Art. Gisscience Journal, 10(5), 1977-1990. DOI: 20.18001.GSJ.2022.V10I5.23.411178
27. Singh, M., Bhalla, A., & Singari, R. M. (2023). Exploring Dadaist Elements in Abstract Designs Inspired by Music for Creating Visually Striking Posters. Gisscience Journal, 10(5), 1991-2004. DOI: 20.18001.GSJ.2022.V10I5.23.411179
28. Singh, M., Singari, R. M., & Bholey, M. (2023). A Review Study of Cognitive Design Research on Colors from a Visual Psychological Perspective. International Journal of Experiment Research and Review, 30, 1-10. DOI: DOI: 10.52756/ijerr.2023.v30.009.
29. Singh, M., Singari, R. M., & Bholey, M. (2023). A Study of Indian Cultural Heritage, Effect of Colors and Human Behavior. ZEICHEN Journal, 109(9), 1-10. DOI: 15.10089.ZJ.2023.V09I02.285311.3074
30. Singh, M., Singari, R. M., & Bholey, M. (2023). Exploring the Role of Colors in Fine Art: A Comparative Study of Traditional and Digital Mediums from Ancient Times to the Modern Era. European Chemical Bulletin, 12, 8725-8759. DOI: 10.48047/ecb/2023.12.si4.781
31. Singh, M., Singari, R. M., & Bholey, M. (2023). Jaipur Miniature Painting: An Exploration of Its Historical Significance, Evolution, and Contemporary Applications in Modern Lifestyle and Product. GIS Science Journal, 10, 1122-1133. DOI: 20.18001.GSJ.2022.V10I3.23.40885
32. Singh, M., Singari, R. M., & Bholey, M. (2023). The Unique Visual and Psychological Effects of Jaipur Blue Pottery: A Study with Cognitive Design Research Approach. Res Militaries, 13, 3488-3505. Retrieved from https://resmilitaris.net/menu-script/index.php/resmilitaris/article/view/3244
33. Singh, M., Singari, R. M., & Maheshwari, B. (2023). Harnessing the Power of Color: The Impact of Color on Visual Psychological Perception in Indian Handicrafts. Delhi Technological University, Delhi, India & Guru Gobind Singh Indraprastha University, Delhi. Journal Name, 9(3), 180-189. doi:15.10089.ZJ.2023.V09I03.285311.3092.
34. Skaggs, R. (1981). Colors can help soothe the paediatrics patient. Hospitals, 55(17), 91.
35. Stumpf, C. (2014). Color psychology: Everything you need to know. AVA Publishing.
36. Umamaheshwari, N., Asokan, S., & Kumaran, T. (2013). Child friendly colors in a paediatrics dental practice. Journal of Indian Society of Pedodontics and Preventive Dentistry, 31(3), 193-196. DOI: 10.4103/0970-4388.121817

The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/2514-4774.htm

Journal of Capital Markets Studies

The rise of passive investing: a systematic literature review applying PRISMA framework

Priya Malhotra
USME East Campus, Delhi Technological University, Delhi, India

Received 3 December 2023
Revised 14 March 2024
23 March 2024
Accepted 8 April 2024

Abstract

Purpose – Passive investing has established itself as the dominant force in the world of professionally managed assets, surpassing the concept of index funds. Its meteoric rise is fueled by investors' preference for its dual benefits of strong diversification and low cost. A comprehensive study of the economic model, addressed areas and market structure has not yet been conducted, despite the existence of numerous studies on more specific topics. To address this gap, this paper examines 943 articles on passive investing published between 1998 and 2022 in SCOPUS and Web of Science.

Design/methodology/approach – The study utilizes the most pertinent tools for conducting a systematic review by the PRISMA framework. This article is the result of SLR and extensive bibliometric analysis. Contextualized systematic literature review is used to screen and select bibliographic data, which is then subjected to a variety of bibliometric analyses. The study provides a bibliometric overview of works on passive investment research that are indexed in Scopus and Web of Science. Bibliometrix, VoS Viewer and Cite Space are the tools used to conduct content and network analysis, to ascertain the present state of research, as well as its focus and direction.

Findings – Our exhaustive analysis yields important findings. One, the previous decade has witnessed a substantial increase in the number of publications and citations; in particular, the inter-disciplinary and international scope of related research has expanded; Second, the top three clusters on "active versus passive funds," "price discovery and market structures" and "exchange-traded funds (ETFs) as an alternative" account for more than fifty percent of the domain's knowledge; Third, "Leveraged ETFs (LETFs)" and "environmental, social and governance (ESG)" are the two emerging themes in the passive investing research. Fourth, despite its many benefits, passive investing is not suitable for everyone. To get the most out of what passive investing has to offer, investors, intermediaries and regulators must all exercise sufficient caution. Our study makes a substantial contribution to the field by conducting a comprehensive bibliometric analysis of the existing literature, highlighting key findings and implications, as well as future research directions.

Research limitations/implications – While the study contributes significantly to the field of knowledge, it has several limitations that must be considered when interpreting its findings and implications. With our emphasis on academic journals, the study analyzed only peer-reviewed journal articles, excluding conference papers, reports and technical articles. While we are confident that our approach resulted in a comprehensive and representative database, our reliance on Elsevier Scopus and Web of Science may have resulted in us overlooking relevant work accessible only through other databases. Additionally, specific bibliometric properties may not be time-stable, and certain common distribution patterns of the passive investing literature may still be developing.

Practical implications – With this study, it has been possible to observe and chart the high growth trajectory of passive investing research globally, especially post-US subprime crisis. Despite the widespread

JEL Classification — C23, G12, G23

© Priya Malhotra. Published in *Journal of Capital Markets Studies*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) license. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this license may be seen at http://creativecommons.org/licences/by/4.0/legalcode

Declaration of conflicting interests: The authors declared no existing or potential conflict of interests with respect to the research, authorship and/ or publication of this paper.

Research involving human participants and/or animals: The author(s) declared that the study was conducted without the involvement of human participants and/or Animals.

Informed consent: The author(s) declare that the study was conducted on well-published secondary data and does not involve any informed consent keeping with research ethics.

Journal of Capital Markets Studies
Emerald Publishing Limited
2514-4774
DOI 10.1108/JCMS-12-2023-0046

adoption of passive investing as an investment strategy, it is not a one-size-fits-all proposition. Market conditions change constantly, and it frequently requires an informed eye to determine when and how much to shift away from active investments and toward passive ones. Currency ETFs enable investors to implement a carry trade strategy in their portfolios; however, as a word of caution, currency stability and liquidity can play a significant role in international ETFs. Similarly, LETFs may be better suited for dynamic strategies and offer less value to a long-term investor. Lastly, the importance of investor education cannot be underestimated in the name of the highly diversified portfolio when using passive alternatives, for which necessary efforts are required by regulators and investors alike.

Social implications – The inexorable trend to passive investing creates numerous issues for fund management, including fee and revenue pressure, which forces traditional managers to seek new revenue streams, such as illiquid and private assets, which also implies increased portfolio risk. Additionally, the increased transparency and efficiency associated with the ETF market indicates that managers must rethink the entire value chain, beginning with technology and the way investments interact. Passive investments have triggered changes in market structure that are still not fully understood or factored in. Active management and a range of valuation opinions on whether a price is "too low" or "too high" provide much-needed depth to a market as it attempts to strike a delicate balance between demand and supply forces, ensuring liquidity at all price points.

Originality/value – I hereby certify that I am the sole author of this paper and that no part of this manuscript has been published or submitted for publication.

Keywords Passive investing, ETF, Index funds, Benchmarking, Bibliometric, Systematic

Paper type Literature review

1. Introduction

Globally, and particularly in developed countries like the United States, there has been a substantial shift in investor preference toward passive investing. Passive asset management, commonly known as passive investing, is an investment approach that monitors a market-weighted index (Sharpe, 1991). It can be accomplished by holding any or a combination of index funds, exchange-traded funds (ETFs), Index future contracts, Index futures options or Stock Market Index Swaps (Hasbrouck, 2003; Stoll and Whaley, 1990; Constantinides *et al.*, 2011) Passive investing, which is most widespread in equities, is gradually becoming more popular in other investment categories such as bonds, commodities and hedge funds (Jaeger and Wagner, 2005). When compared to actively managed funds, it provides good diversification, minimal transaction costs and lower turnover (Baks *et al.*, 2001; Pézier and White, 2008).

Passive investing is founded on the efficient market hypothesis (EMH), which was developed in the 1960s and contends that shares always trade at their fair market value and that persistent alpha generation is impossible. This theory rejected the efficacy of stock selection and market timing and represented the earliest school of thought on passive investing. Index funds were introduced in the 1970s and have been restricted to large pension and social security institutions since then (Sorensen *et al.*, 1998; Clark and Hebb, 2004; Elton *et al.*, 2007; Ayres and Curtis, 2014). Gruber (1996) brought attention to the paradox that investors purchase actively managed funds even though they underperform index funds on average. With growing investor dissatisfaction with active fund underperformance (Sorensen *et al.*, 1998; Malkiel, 2005; Sorensen *et al.*, 1998) and a desire to reduce costs across social benefit schemes, the relative appeal of passive investments has increased by multiple degrees in the aftermath of the 2008–2012 subprime crisis (Elton *et al.*, 2002). This has piqued the interest of industry, academia, and researchers alike, resulting in a slew of studies examining their performance (Avellaneda and Lee, 2010). The literature on passive investing, hence, has exploded in recent years.

The testing of the EMH, the comparative performance assessment of passive portfolios to curated folios, and investor biases in passive investing have all been topics of discussion in the past (Frino and Gallagher, 2001; Elton *et al.*, 2004; Choi *et al.*, 2010; Guercio and Reuter, 2014; Cremers *et al.*, 2016; Fahling *et al.*, 2019). Multiple studies addressing various characteristics of index funds, such as fee expense (Hortacsu and Syverson, 2004; Elton *et al.*, 2012), return

volatility (Goetzmann and Massa, 2003), tracking error (TE) (Meade and Salkin, 1990; Chen *et al.*, 2006), liquidity, investor purchase and redemption patterns (O'Neal, 2004), price discovery (Chen and Gau, 2009), hidden costs (Petajisto, 2011) and portfolio allocation (Fan *et al.*, 2016) have been conducted since the inception of index funds. Recent research on index funds has progressed to sparse portfolios (Takeda *et al.*, 2013), goal programming (Wu *et al.*, 2007), algorithm-assisted allocations (Schoenfeld, 2017) and financial innovations (Cremers *et al.*, 2016).

Lo (2007) integrated the EMH with behavioral finance to create the adaptive market hypothesis (AMH), which states that rational investors can swiftly become irrational in reaction to increased market volatility. In the aftermath of the US Subprime crisis, index funds were heavily criticized for causing the 2007–2008 spike in commodity futures prices, consequently, numerous studies examined the effect of index funds on the prices of underlying assets (Malkiel and Radisich, 2001; Michał Falkowski, 2011; Irwin and Sanders 2011; Ivanov, 2013; Hamilton and Wu, 2015).

ETFs, which manage trillions of dollars (USD) in capital flows in developed economies, are the most recent and significant development in modern capital markets after index funds. Since their introduction, various barometers of ETF performance have been explored in depth, including TE (Johnson, 2009; Shin and Soydemir, 2010; Chu, 2011), pricing efficiency (Israeli *et al.*, 2017) and fund flows (Clifford *et al.*, 2014). As ETFs gained prominence, their role in price discovery and information transmission (Tse and Martinez, 2007; Levy and Lieberman, 2013; Haslam *et al.*, 2018), volatility in prices of underlying (Gleason *et al.*, 2004), market liquidity (Boehmer and Boehmer, 2003; Ben-David *et al.*, 2017), market microstructures (Gao *et al.*, 2018), market co-integration and error correction (Chou and Chung, 2006; Delcoure and Zhong, 2007; Broman, 2016), volatility spillover (Krause and Tse, 2013; Inci, 2018) has been investigated in detail. With growing environmental consciousness, ETF research has shifted to evaluating the efficacy of investments in alternative energy (Miralles-Quirós and Miralles-Quirós, 2019). Furthermore, inter se comparison between index funds and ETFs on a variety of criteria, including substitutability, cost structure, investor types and risk-adjusted returns have been conducted (Beneish and Gardner, 1995; Kostovetsky, 2003; Harper *et al.*, 2006). Leveraged ETFs (LETFs) are one of the most recent financial developments in the ETF realm, and various studies have been conducted to explain this concept (Avellaneda and Zhang, 2010; Charupat and Miu, 2011).

Notwithstanding economic challenges, unpredictability and an initial upheaval precipitated by the coronavirus disease 2019 (COVID-19) pandemic, the asset management sector rebounded in 2022 and produced favorable returns. The assets under management (AUM) of passive funds are projected to reach USD 34 trillion by 2025, according to the Global Asset Management Report, (2021) by Boston Consulting Group. Systematic reviews serve the purpose of identifying particular domains that necessitate further research. Additionally, they enable professionals to obtain a comprehensive understanding (Randhawa *et al.*, 2016). Previous studies have conducted literature reviews on distinct aspects of passive investing strategies, such as fixed-income and index investing (Tang and Xu, 2017), the effects, trading, and institutions of ETFs (Madhavan, 2012; Joshi and Dash, 2024), and an overview of Indian index funds (Sarkar *et al.*, 2013). Despite the expanding corpus of scholarly work concerning specific facets, a comprehensive overview of the overarching passive investing strategy remains extremely scarce. Our research is most closely aligned with the findings of the Fisch *et al.* (2019) study. In light of the market mechanism, the study offers a comprehensive analysis of the theoretical framework from institutional, policy and regulatory standpoints to enlighten passive investors. However, a review of the existing literature in this field is not included in the study. Given these overarching concerns, the expansion and development of the corpus of research in numerous disciplines emphasizes the need for a quantitative investigation to establish a comprehensive framework, identify noteworthy trends and determine directions for future research. Our study investigates the subsequent research

inquiries by employing a systematic literature review (SLR) and a rigorous bibliometric analysis.

RQ1. What characterizes the current research trend in passive investing?

RQ2. Which sources, articles and authors are the most influential in this research domain?

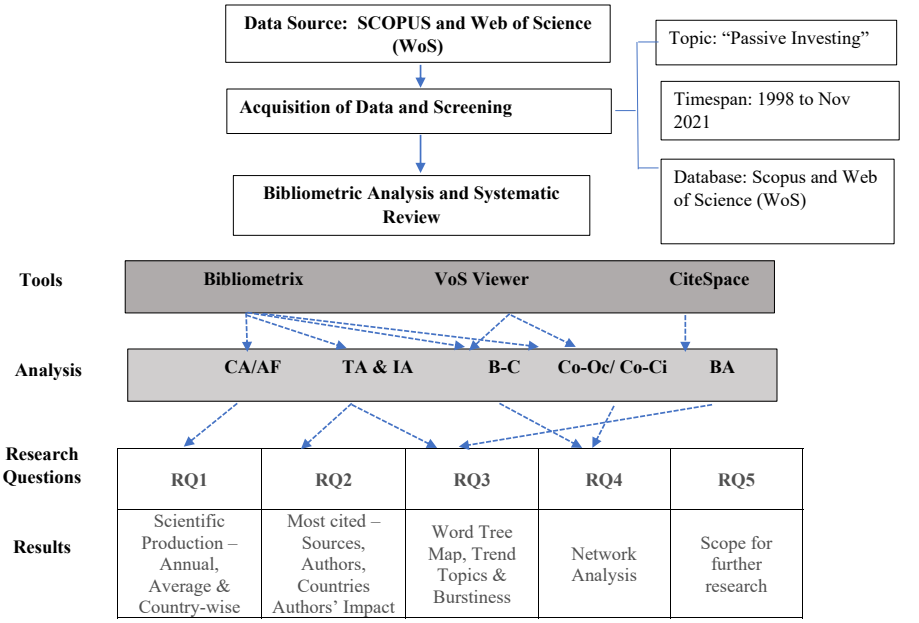
RQ3. What are the dominating themes in this field of study?

RQ4. What has been the evolution of the conceptual and intellectual structure (IS) over time?

RQ5. What is the future research scope?

2. Data and research methodology

This article is the result of SLR and extensive bibliometric analysis. Contextualized SLR (Tranfield *et al.*, 2003) is used to screen and select bibliographic data, which is then subjected to a variety of bibliometric analyses (Zupic and Čater, 2015). Bibliometric analysis as a technique entails the application of statistical tools to the examination of the bibliographic attributes of a body of literature. We follow Pattnaik *et al.* (2020) as a methodical approach for data selection and research design as depicted in Figure 1.



Note(s): Abbreviations utilized: CA/AF – Citation Analysis/ Analysis of Frequency; IA – Impact Factor Analysis; B-C – Bibliographic Coupling; Co-Oc – Co-Occurrence of keywords/titles/abstracts; Co-Ci- Co-citation analysis of References/Sources/Authors; BA- Burstiness Analysis of Citation/ References/ Sources

Source(s): Authors' own work

Figure 1.
Data selection and research design

2.1 Search criteria and data retrieval

The collected data adhered to a systematic research methodology that outlined the essential stages required to perform a literature review following the PRISMA framework. For the presentation of evidence in systematic reviews and meta-analyses, PRISMA is the acknowledged standard. Journals that underwent peer review were the only sources considered for this study. The evaluation process excluded reports, conference proceedings, publications and theses. Analyses of 943 articles indexed by Scopus and WoS were conducted to address the aforementioned research questions. In selecting these documents, five criteria were considered: (1) Refereed academic journals (to ensure the highest quality of refereeing), (2) Articles published between 1998 and 2021, (3) Indexed by Scopus and WoS, (4) Filtered using keywords and Boolean operators: "passive investing or passive invest*," or "index fund*," or "exchange-traded fund" or "ETF*" or "indexing,"; (5) The study includes all eligible non-duplicate publications in the English language that met the filtration criteria. The process of retrieving meta-data is summarized in Table 1.

2.2 Research methodology

Bibliometric approaches or analyses are now firmly established as an essential component of research evaluation methodology, particularly in scientific and applied disciplines (Ellegaard and Wallin, 2015). It helps to eliminate subjective bias in qualitative literature synthesis (Nerur *et al.*, 2008) and has been widely used in the past to analyze massive amounts of bibliographic content on topics (Blanco-Mesa *et al.*, 2017), universities (Merigó *et al.*, 2019), countries (Mas-Tur *et al.*, 2019) and journals (Donthu *et al.*, 2020). Citation and co-citation analysis are two of the most commonly used bibliographic approaches for uncovering differences in the intellectual landscapes of cited sources (Kessler, 1963; Small, 1973).

Bibliometrix, VoS Viewer and Cite Space are the tools used to conduct content and network analysis and to ascertain the present state of research in this sector, as well as its focus and direction. The descriptive analysis summarizes the trend in publications and citations (Tsay, 2009). The most influential sources, articles and authors are selected based on *h-index*, *g-index* and *m-index* (Hirsch, 2005; Ding and Cronin, 2011). Keyword analysis is used to elucidate the study domain's knowledge framework, emerging patterns and evolving themes (Hu *et al.*, 2018). The conceptual structure is deduced using word-tree analysis (WTM) (Callon *et al.*, 1983). Network analysis is performed using the bibliographic coupling (B-C) principle and network centrality strength to uncover interconnected communities (Bastian *et al.*, 2009).

Steps	Filtration criteria	Excluded	Accepted
Identification	Record identified through database screening (Boolean Operator)		1,320
Screening	Filtration on the basis of the subject (Business, Economics, Econometrics, Finance, Management, Accounting and Social Science)	308	1,012
	Duplicates removed	5	1,007
Eligibility	Document Type (Article and Review)	14	993
	Publication Stage (Final)	10	983
	Language criteria (English)	20	963
Included	Based on the content	20	943

Note(s): No restrictions on the ranking of Journals were made for meta-data selection
Source(s): Authors' own work

Table 1.
PRISMA flow diagram

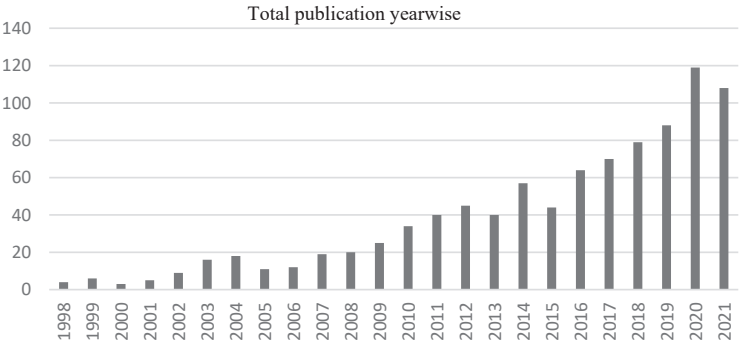
3. Descriptive statistics

3.1 Trend in research and scientific productivity

Addressing RQ1 on the research trend in passive investment, Figure 2 shows that the number of annual publications has been increasing since 2007. To be specific, over 90% of the literature included in this study was published after 2008, implying that the majority of intellectual debates in the area of passive investment occurred in the aftermath of the 2008–2012 global economic crisis.

3.2 Most influential authors, documents and sources

This section addresses the RQ2 concerning the prominent authors, articles, and sources in the field of Passive Investing. Table 2 summarizes the most influential authors on passive investing, arranged by the number of publications; additional metrics are included in the table. Tse Y is ranked first with 14 articles between 1998 and 2001, an h-index of 9, and 241 citations. Gruber MJ is the most prominent author on passive investing, with a TC count of 1,520, followed by Irwin SH (TC: 776) and Elton EJ (TC: 667). The h-index metric measures both citation and productivity impact at the author level; Irwin SH, with an h-index of 10, g (12) and m (0.833), is the most influential author in the field of passive investment, followed by Tse Y (9, 14 and 0.474).



Source(s): Authors' own work

Figure 2.
Annual scientific production

Author	*H*	*g*	*m*	TC	NP
Tse, Y.	9	14	0.474	241	14
Irwin, S. H.	10	12	0.833	776	12
Gruber, M. J.	8	11	0.242	1,520	11
Elton, E. J.	7	10	0.212	667	10
Sanders, D. R.	8	9	0.667	643	9
Frino, A.	6	8	0.286	178	8
Rompotis, G. G.	5	8	0.357	72	8
Ivanov, S. I.	4	7	0.364	80	7
Lee, C. C.	3	6	0.429	38	7
Xu, L.	3	5	0.6	32	7

Note(s): The table above enlists the top 10 authors contributing to the extant literature based on three distinct impact factors: h-index, g-index and m-index (*h*, *g* and *m* respectively). It also presents the total citation count (TC) accredited to the author along with the number of papers (NP) by the author

Source(s): Authors' own work

Table 2.
Most influential authors on passive investing based on three impact factors

As assessed by NP, Table 3 summarizes the top journals in the field of passive investment. The *Journal of Portfolio Management* tops the list with 35 contributions to the current body of knowledge, followed by Managerial Finance (NP: 28) and the *Journal of Banking and Finance* (NP: 25). It's worth noting that the top ten journals account for more than a quarter (26%) of the papers examined in this study, while the top twenty journals account for almost 40% of the total articles examined. While the Journal of Banking and Finance is the most widely cited source in the domain of passive investment, according to the *h* and *g* indexes (14 and 23 respectively), the Journal of Finance (not included in the table below) ranks first with 1,684 citations using TC as a measure of influence. Publications in reputable journals demonstrate the study domain's growth, inspiring additional research endeavors (Acedo *et al.*, 2006). The aforementioned considerations, hence, justify the remarkable increase in research in this area.

Based on the citation score, Table 4 presents the most influential papers and references on passive investing. Local citations (LCs) are based on a dataset included in the study, whereas global citations (GCs) are based on a large dataset available on Google Scholar and Web of Science. The top most influential article (GC:853) is Gruber J.'s "*Another Puzzle: The Growth in Actively Managed Mutual Funds*" (1996), followed by Shleifer A.'s "*Do Demand Curves for Stocks Slope Down?*" (1986) (GC: 589). Using LC as a metric of influence, Elton J., "*Spiders: Where Are the Bugs?*" (2002) ranks first (LC:71), followed by GruberJ., "*Managed Mutual Funds*" and Ben-David *et al.* (2018) "*Exchange-Traded Funds*" (LC: 49).

3.3 Most dominating themes in research on passive investing

In response to RQ3, Figure 3 offers an overview of the thematic evolution seen in the literature on passive investing. The frequency with which keywords appear in the literature from 1998 to the current day is used to extract the most prominent themes. Time slices were created using significant economic events that affected the development of financial markets in

Source	*H*	*g*	*m*	TC	NP
Journal of Portfolio Management	12	21	0.500	498	35
Managerial Finance	7	9	0.269	125	28
Journal of Banking and Finance	14	23	0.467	555	25
Journal of Asset Management	6	11	0.462	137	19
Journal of Index Investing	1	1	0.250	5	4
International Review of Financial Analysis	7	11	0.500	129	15
Investment Management and Financial Innovations	4	7	0.250	58	11
Journal of Futures Markets	7	13	0.368	184	15
Financial Analysts Journal	9	16	0.346	311	16
Journal of Alternative Investments	3	6	0.231	120	6
Applied Economics	5	9	0.263	94	9
Applied Economics Letters	5	7	0.417	63	11
Financial Review	8	11	0.444	210	11
Journal of Empirical Finance	7	10	0.389	366	10
Quantitative Finance	4	10	0.333	156	10
North American Journal of Economics and Finance	5	8	0.625	69	11
Review of Financial Studies	7	11	0.412	376	11
Journal of Financial Markets	8	11	0.471	261	11
Journal of Wealth Management	3	4	0.273	22	7
Applied Financial Economics	6	10	0.300	129	10

Note(s): The table above enlists the top 20 sources ranked in the order of number of papers (NP) on passive investing along with total citations (TC). The source's impact factor measured as *h*-, *g*- and *m*-index are also mentioned in the table

Source(s): Authors' own work

Table 3. Most influential sources

Author	Title	Year	LC	GC	LC/GC ratio (%)
Elton *et al.* (2002)	"Spiders: Where are the bugs?"	2002	71	105	67.62
Gruber (1996)	Another Puzzle: The Growth in Actively Managed Mutual Funds	1996	49	853	5.74
Ben-David *et al* (2017)	Exchange-Traded Funds	2017	49	41	119.51
Ben-David *et al.* (2018)	Do ETFs Increase Volatility	2018	43	82	52.44
Hasbrouck (2003)	Intraday Price Formation in US Equity Index Markets	2003	39	229	17.03
Shleifer (1986)	Do Demand Curves for Stocks Slope Down?	1986	37	589	6.28
Delcoure and Zhong (2007)	On the premiums of iShares	2007	36	52	69.23
Tse and Martinez (2007)	Price discovery and informational efficiency of international iShares funds	2007	31	52	59.62
Poterba and Shoven (2002)	Exchange-Traded Funds: A New Investment Option for Taxable Investors	2002	30	71	42.25

Table 4. Most influential articles

Note(s): Table above enlists the top 10 cited documents on passive investing along with the local citations (LC) and Global citations (GC), author(s) and year of publication for the respective documents
Source(s): Authors' own work

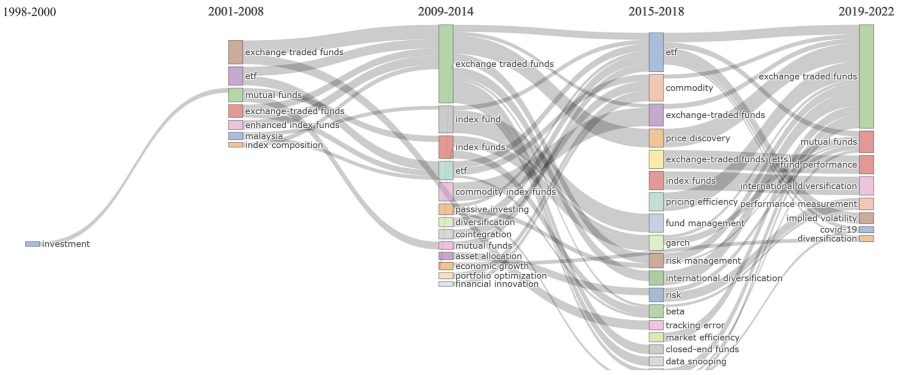


Figure 3. Thematic evolution in passive investing research

Note(s): Figure 3 presents the thematic evolution of research in the area of passive investing across four time slices, the frequency of occurrence is the qualifying criteria for the ranking of themes
Source(s): Authors' own work

general and passive investments in particular. The first time slice, from 1998 to 2000, illustrates the closed character of financial investments before the US IT boom. With the turn of the twenty-first century and the integration of financial markets, *exchange-traded funds* and *enhanced index funds*, which seek to boost the returns produced by strict index funds, were added to the broader basket of passive investments, which index funds previously dominated. The rapid post-global financial crisis (GFC) economic recovery of the Malaysian economy [1], bolstered by robust government spending and the strengthening of foreign trade, attracted the research community's attention, as evidenced by its inclusion in the prominent themes in the period spanning 2001–2008. Passive investments experienced

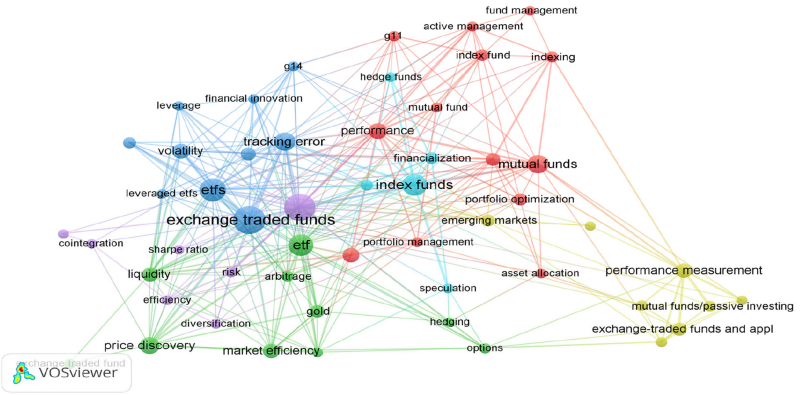
substantial growth in the aftermath of the GFC of 2007–2008, gaining extensive academic interest with the addition of *commodity index funds*, *financial innovations* and *asset allocation* to its repertoire. The increased availability of soft liquidity after quantitative easing (QE) has resulted in developing economies emerging as favored investment destinations. This development provides further rationale for the inclusion of *diversification* in the most prolific themes, alongside *portfolio optimization*. Additionally, heightened risk management awareness is a consequence of the GFC; as a result, international investors have redirected their attention toward attaining *international diversification* via ETFs that replicate global indices. The aforementioned developments have also prompted the incorporation of *TE* and *pricing efficiency* metrics into ETF performance evaluations, the implementation of *data snooping* and *GARCH* to identify patterns and the rise of leveraged ETFs (LETFs) as a novel concept in passive investing from 2015 to 2018. In 2019, asset markets experienced significant disruption and dislocation following the outbreak of COVID-19, the deadliest threat of the century. The fourth time slice commenced in that year. As a consequence, investigations into *implied volatility* in *performance evaluation, price discovery, portfolio construction* and other ETF-related studies were augmented to incorporate the quintessential *COVID-19* perspective. As environmental and sustainability concerns have grown, environmental, social and governance (ESG) orientation has emerged as a prevalent topic in passive investment research and a popular catchphrase.

4. Empirical findings

This section elaborates on the conceptual and intellectual framework of the existing literature concerning RQ4. It accomplishes this by employing a range of methodologies, including keyword analysis, B-C, keyword co-occurrence and content analysis, to identify key clusters of scholarly work in the domain of passive investing.

4.1 Network visualization of the literature

4.1.1 Keyword network analysis. In the form of a keyword co-occurrence network of the top 53 terms, Figure 4 depicts the conceptual structure of the scientific literature under investigation. Six clusters make up the network. The top cluster has 25% of all frequently co-occurring keywords in the literature, whereas the second and third clusters have 19% of all co-occurring keywords. While the fourth, fifth and sixth positions, respectively, contain 15, 13



Source(s): Authors' own work

Figure 4.
Keywords co-occurrence network

and 9% keywords. Each keyword is a node, the size of which is determined by its frequency of occurrence. The terms "exchange-traded funds," "tracking error," "index funds," "mutual funds," "performance measurement," "exchange-traded funds" and "etf" dominate in terms of network strength, implying the most prolific empirical themes in passive investing research.

4.1.2 Bibliometric content analysis of the network. Figure 5 shows the network illustrating the B-C of the top 251 articles. The network comprises nine clusters, each of which contains articles with a minimum of ten citations. The first three clusters represent more than half of all articles examined for the B-C, whereas the first five clusters represent more than 70% of all articles. The top three most popular articles in each cluster are listed in Table 5, sorted by page rank (Ding, 2011).

4.1.3 Clusterwise content analysis. With the arrival of the Science Citation Index (SCI) on the bibliographic landscape, the technique of analyzing vast scientific literature on the basis of the relationship between published papers (measured by co-referencing) has garnered huge research interest (Martyn, 1964). The B-C combined with the cluster method could provide valid depictions of the interrelationship between articles with similar referencing (Kessler, 1963). This section gives a clusterwise depiction of research in passive investing.

Cluster 1: active funds vs passive funds

The first cluster consists of 44 studies conducted from 1996 to 2019. The cluster's overarching theme is the comparative performance assessment of active funds with index funds and other passive benchmarks. The cluster is founded on the work of Gruber (1996), a research that compared active and passive funds and casts doubt on active mutual funds' meteoric expansion despite delivering lackluster performance on average (Moore *et al.*, 1999). The empirical inquiry is complemented by Elton *et al.* (1996)'s discussion of the persistence and predictability of risk-adjusted returns and fund flows as also how modern portfolio theory may be used to construct optimal portfolios generating economically significant expense-adjusted returns. The performance differential between active and passive funds (primarily index funds) has been examined from a variety of perspectives, including portfolio allocations (Keim, 1999; Baks *et al.*, 2001; Ahmed and Nanda, 2005; Bollen and Busse, 2006; Gil-Bazo and Ruiz-Verdu, 2009; Choi *et al.*, 2010; Cuthbertson *et al.*, 2010) and, non-portfolio differentiators such as fund fee expense, cash flows, trading costs and the fund management (Baks *et al.*, 2001; Gastineau, 2004; Golec, 2003; Hortacsu and Sveyorson, 2004; O'Neal, 2004; Houge and Wellman, 2007; Rompotis, 2009; Elton *et al.*, 2012; Bogle, 2014; Cremers *et al.*, 2016). Numerous studies have investigated anomalies such as in fund flows (Białkowski *et al.*, 2013) and fund

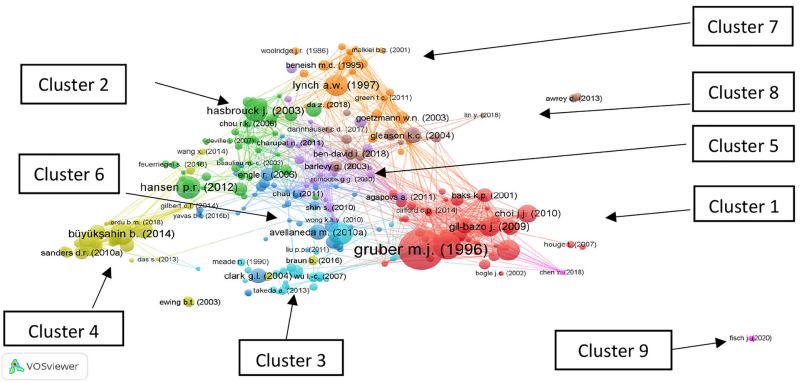


Figure 5.
Bibliographic coupling of documents (by number of citations)

Source(s): Authors' own work

Author	Title	Cluster	TC	PY	PR
Hortacsu and Syverson (2004)	Product differentiation, search costs and competition in the mutual fund industry: a case study of s&p 500 index funds	1	217	2004	1.44
Gruber (1996)	Another puzzle: the growth in actively managed mutual funds	1	853	1996	1.41
Elton *et al.* (1996)	The persistence of risk-adjusted mutual fund performance	1	336	1996	0.81
Hasbrouck (2003)	Intraday price formation in US equity index markets	2	229	2003	2.1
Elton *et al.* (2002)	Spiders: where are the bugs?	2	105	2002	1.98
Hansen *et al.* (2012)	Realized Garch: a joint model for returns and realized measures of volatility	2	265	2012	0.58
Clark and Hebb (2004)	Pension fund corporate engagement: the fifth stage of capitalism	3	112	2004	0.55
Avellaneda and Lee (2010)	Statistical arbitrage in the us equities market	3	123	2010	0.31
Malkiel and Radisich (2001)	Reflections on the EMH: 30 years later	3	93	2005	0.29
Irwin and Sanders (2012)	Testing the masters' hypothesis in commodity futures markets	4	152	2012	1.52
Büyükşahin and Robe (2014)	Speculators, commodities and cross-market linkages	4	229	2014	0.97
Irwin and Sanders (2011)	Index funds, financialization and commodity futures markets	4	211	2011	0.94
Madhavan (2012)	Exchange-traded funds, market structure and the flash crash	5	63	2012	0.91
Charupat and Mu (2011)	The pricing and performance of leveraged exchange-traded funds	5	52	2011	0.68
Agapova (2011)	Conventional mutual index funds versus exchange-traded funds	5	62	2011	0.6
Gasbarro *et al.* (2007)	Stochastic dominance analysis of ishares	6	48	2007	1.22
Fong *et al.* (2005)	International momentum strategies: a stochastic dominance approach	6	92	2005	1.07
Ippolito (1989)	Efficiency with costly information: a study of mutual fund performance, 1965–1984	6	255	1989	0.6
Goetzmann and Massa (2003)	Index funds and stock market growth	7	79	2003	1.51
Lynch and Mandenhall (1997)	New evidence on stock price effects associated with changes in the s&p 500 index	7	232	1997	0.92
Beneish and Gardner (1995)	Information costs and liquidity effects from changes in the dow Jones industrial average list	7	59	1995	0.39
Ben-David *et al.* (2018)	Do etfs increase volatility?	8	82	2018	4.34
Goetzmann and Massa (2002)	Daily momentum and contrarian behavior of index fund investors	8	70	2002	0.47
Gleason *et al.* (2004)	Analysis of intraday herding behavior among the sector etfs	8	129	2004	0.32
Chen and Scholtens (2018)	The urge to act: a comparison of active and passive socially responsible investment funds in the United States	9	13	2018	1.87
Fisch *et al.* (2019)	The new titans of wall street: a theoretical framework for passive investors	9	14	2020	1.55
Bebchuk and Hirst (2020)	The specter of the giant three	9	13	2019	1.24

Note(s): The table lists the most prolific top 3 articles of the nine bibliographic clusters created on the basis of total citations (TC). PageRanks (PR) are used to rank the articles. Publishing year (PY) and First Author name are also mentioned in the table

Source(s): Authors' own work

Table 5.
Most prolific articles on passive investing

manager names (Kumar *et al.*, 2015). Brown *et al.* (2007) and distribution channels found no conclusive evidence of underperformance relative to index funds (Guercio and Reuter, 2014). The cluster also covers quintessential comparative performance assessment studies. Lettau and Madhavan (2018) conducted an in-depth examination of the structure and organization of ETFs and various advantages offered by them with a brief reference to the disadvantages arising from market frictions. Kostovetsky (2003) discovered that ETFs are more suitable for long-term investors and that flows to ETFs are determined by individual investor characteristics and risk tolerance (Clifford *et al.*, 2014; Mikhaylov *et al.*, 2019). Deviations from the underlying benchmark, TE, its determinants and persistence have been the most researched dimension in ETF studies (Frino *et al.*, 2004; Gastineau, 2004). Blitz *et al.* (2012) expanded the analysis to include Emerging Market Economy (EME)-focused ETFs and reported much higher TE in ETFs with a developed economy focus. Notably, ETFs deliver on their underlying objective of tracking indexes and the addition of active components adds no value to ETFs Sherrill *et al.* (2017).

Cluster 2: effect of market structures

The second cluster consists of 43 publications published between 2002 and 2019. The cluster's major focus is *price formation, price discovery, market structure, and its effect on volatility and liquidity*, with ETFs as the dominant investment type. Except during periods of extreme volatility, the contribution of ETFs tracking the broad market is considerable in a complex financial market environment comprised of equity indexes, ETFs and micro future contracts (Hasbrouck, 2003; Tse and Erenburg, 2003; Chen *et al.*, 2016; Chen and Tsai, 2017). The role of spot market dynamics in comparison to futures markets has been extensively studied, and it has been determined that the introduction of small tick sizes (Beaulieu *et al.*, 2003; Wu *et al.*, 2007; Chen and Gau, 2009), market microstructure and fragmentation (Hendershott and Jones, 2005; Tse *et al.*, 2006) are economically significant. Before the introduction of computerized trading, the futures markets dominated the process of price discovery (Schlusche, 2009). Notably, ETFs now enabled by in-kind creation and redemption activity in the stock market are now leading price discovery in commodity markets (Ivanov, 2013). Increasing institutional ownership and algo-trading using high-frequency data is further aiding the price discovery mechanism (Elton *et al.*, 2002; Charupat and Miu, 2011; Chen *et al.*, 2016; Nugroho, 2021). Volatility measurement both before and after the introduction of ETFs (Hansen *et al.*, 2012) has been another frequently researched dimension, as has market fragmentation. Volatility-return relationship (Daigler *et al.*, 2014), volatility spillover and leverage effects on returns (Chen and Huang, 2010; Krause and Tse, 2013; Dedi and Yavas., 2016), as well as return predictability (Shynkevich, 2016) have been the subject of research under numerous studies. While modeling ETF returns, multiple realized volatility measures outperform single realized volatility measures, and when multiple forecasting approaches are compared, algorithmic methods and Bagging (Bootstrap aggregation) are found to be more accurate (Brownlees *et al.*, 2011). The impact of the advent of these modern trading platforms on price variations, price formation and liquidity in the ETF market, particularly with an EME concentration, remains an open subject of intellectual dispute.

Cluster 3: ETF as an investment alternative

The third cluster contains 41 publications published during 17 years beginning in 2002. The cluster's major issue is the *ETFs as an investment alternative*, with a particular emphasis on trading strategies for passive investing. ETFs combine the characteristics of closed-end funds (CEFs) and open-end funds (OEFs) (Harper *et al.*, 2006). While CEFs and ETFs are similar in terms of trading and organization, liquidity and arbitrageability set these apart (Hughen and Mathew, 2009). Both fund types overreact to domestic results; however, ETFs

more closely track the underlying index (Aber *et al.*, 2009). The gap between the net asset value (NAV) and ETF price, indicating pricing efficiency can be explained by measurement errors. However, the rate of error correction is faster in domestic ETFs compared to international ETFs with implications on portfolio diversification (Jares and Lavin, 2004; Engle and Sarkar, 2006). Price deviation (premium/discounts) has been a subject of continual investigation in ETF research; ETFs are more likely to trade on a premium basis than on a discount basis. Delcoure and Zhong (2007) argue that factors such as institutional ownership, bid-ask spread, trading volume, currency volatility, geopolitical crisis and conditional correlation between markets contribute to the formation of iShare premiums. Trading methods based on passive investments are a major focus of research in this cluster. Avellaneda and Lee (2010) examined the efficacy of model-driven statistical arbitrage and determined that strategies based on exchange-traded funds yielded a higher Sharpe ratio. Cross-border spillovers (Lee and Chen, 2020), serial dependence in returns (Baltussen *et al.*, 2019), feedback trading (Chau *et al.*, 2011) and the use of ETFs to capitalize on industry and country momentum have all been examined to determine the international diversification benefits that passive investments may provide. While ETFs have a reduced average bid-ask spread (Andreu *et al.*, 2013), index funds provide greater diversification benefits due to their lower correlations with the underlying index (Khorana *et al.*, 1998). Individual investors can easily include carry-trading into their asset allocation strategy through the use of ETFs (Das *et al.*, 2013). Gold ETFs and other index funds tracking Gold offer hedging and safe-haven features (Pullen *et al.*, 2014). Leveraged exchange-traded funds (LETFs) are the most sophisticated ETF concept. However, when Gold is used as the underlying, these exhibit tracking incapability (Leung and Ward, 2015). By and large, LETFs underperform static strategies and are regarded as unsuitable for long-term investors (Avellaneda and Zhang, 2010). The role of behavioral characteristics or investor sentiment in ETF and index fund mispricing presents future research opportunities.

Cluster 4: pricing and market efficiency

The fourth cluster consists of 28 studies that span 16 years, from 2002 to 2018. With commodities at the center of the argument, research in this cluster examines the *impact of passive investment on future prices, risk premia, market efficiency* and EMH testing. Index funds have long been blamed for increasing volatility, distorting price relationships and fueling asset commodity price bubbles (Sanders *et al.*, 2010). Empirical studies examining the relationship between index fund positions and commodity futures prices, however, found no conclusive evidence of index fund positions' impact on returns and volatility (Capelle-Blancard and Couibaly, 2011; Stoll and Whaley, 2011; Hamilton and Wu, 2015), Büyükşahin and Robe (2014) reported similar results during periods of market stress. A novel study by Gilbert and Pfuderer (2014) provided greater evidence that changes in index positions offer prognostic information for commodity prices. The financialization of commodities results in the blurring of the distinctions between commodities and stock markets (Michał Falkowski, 2011). Ordu *et al.* (2018) cited variables such as currency appreciation, availability of liquidity and the expansion of developing markets contribute to the increased volatility in raw and soft commodities markets. This problem has also been discussed with the risk premia accessible to commodity investors. Main *et al.* (2018) assert that before and after financialization, the average unconditional return on individual commodity futures markets is equal to zero.

Cluster 5: market regimes and pricing

The fifth cluster has 26 articles spanning eight years from 2010 to 2020. This cluster's central theme is *market structure, market niches and the effect of various market regimes* on the performance and pricing efficiency of passive investments. The 2010 flash crash served as a

proving ground for a variety of economic theories. Madhavan (2012) revealed that market fragmentation was a significant factor in the propagation of the liquidity shock and how securities were affected by the flash crash. Price deviations and mispricing in ETFs have been thoroughly researched since then across a range of market environments. While Wong and Shum (2010) concluded that ETF returns are not proportionate to volatility, Shanmughama and Zabiulla (2012) discovered a substantial variation in alpha generation between bearish and bullish markets, with TE and average premium being larger in bearish conditions and vice versa. As markets become more integrated, the changing market dynamics have been a continuing source of inquiry. Petajisto (2017) found that funds that invest in overseas or illiquid equities experience more price variations. Ackert and Tian (2008) assert that when country ETFs' fund premiums are compared to US ETFs' fund premiums, country ETFs demonstrate a high degree of positive autocorrelation. Their undervaluation is a result of scale effects, momentum, and illiquidity. In the Asian markets context, Shin and Soydemir (2010) found greater persistence in TEs, with currency rates playing a substantial role in the failure to track along with commonly cited market liquidity factor (Osterhoff and Kaserer, 2016; Camilleri *et al.*, 2020). On the other hand, Charteris *et al.* (2014) provide compelling evidence that ETF premiums and discounts encourage feedback trading in EMEs, which becomes more prevalent as premiums rise. LETF is a derivative-based exchange-traded fund that seeks daily returns that are a multiple of the underlying index's return. According to Tang and Xu (2013), market frictions and inefficiency greatly contribute to the LETF's return deviation. On the other hand, Li and Zhao (2014) examined how LETF trading affects the market quality of component stocks and reported a positive correlation between the trading volumes of the LETF and the stocks. ETFs have attracted enormous investor interest, making them one of the most significant financial innovations. Empirical research has discovered a strong correlation between Information and Communications Technology (ICT) penetration and ETF popularity and adoption as an investment class. The dark side of financial innovation in the domain of high-frequency data, algorithmic trading and information diffusion, on the other hand, remains a relatively untapped field of study.

Cluster 6: portfolio strategies

This cluster includes 26 studies that examine portfolio construction and optimization methodologies for passive investing. The cluster dates back more than three decades, with Ippolito (1989) providing the first evidence of the efficiency of costly information in mutual fund returns consistent with optimal trading in efficient markets. Numerous studies found that portfolios constructed using stock-specific criteria (Meade and Salkin, 1990), diversity-weighted index (Fernholz *et al.*, 1998), enhanced index investing (Wu *et al.*, 2007) and heuristics (Takeda *et al.*, 2013) generated higher returns than portfolios constructed using the market capitalization method. Estrada (2008) demonstrated that low-cost value-based strategies outperformed both price-insensitive passive and capitalization-based indexes in global portfolio building. Thompson *et al.* (2006) conducted exploratory data analysis and found data-driven programming to be superior to the several fundamental models of neoclassical computational finance. In the ETF context, numerous studies found that ETFs provide fewer diversification benefits than implied by the benchmark index's return characteristics (Bhattacharya *et al.*, 2017; Buetow and Henderson, 2012; Pedersen, 2018). Huang and Lin (2011), on the other hand, concluded that ETFs have a higher Sharpe ratio than the target market index, particularly in EMEs. With growing awareness of the interdependence of ESG challenges, sustainable investment has gained traction, allowing investors to benefit from investing in ETFs tracking firms with sustainable development goals (Miralles-Quirós and Miralles-Quirós, 2019). Analyzing hedge strategies based on

restricted passive portfolios, smart beta strategies, alternative energy and ESG investment in general gives further scope for research.

Cluster 7: varying composition of underlying indices

A total of 24 studies examining *changes in the composition of broad market indices* such as the S&P 500, the Russell 2000, the FTSE100, and the DJIA, as well as the associated price effects, form the heart of this cluster, which spans over 2 decades (1994–2018). Index funds, ETFs and other passive funds collectively represent a sizable and important institutional investor class for stocks trading on global stock markets. Any change in the composition of the larger indices tracked by index funds and ETFs causes rebalancing of index fund portfolios and so affects returns (Shleifer, 1986; Lynch and Mendelhall, 1997; Chen *et al.*, 2006), providing vital insight into market efficiency (Madhavan, 2003). This effect could be transient, consistent with the price pressure hypothesis (Woolridge and Ghosh, 1986; Vespro, 2006), or it could be permanent, consistent with Scholes's (1972) price shift followed by considerable abnormal trading volume, confirming downward-sloping demand curve hypothesis (Biktimirov, 2004). Beneish and Gardner (1995) suggested that information symbolism, price pressure, imperfect alternatives and information cost could all be explanatory variables for the stock market consequences of DJIA revisions. On the other hand, Malkiel and Radisich (2001) find no evidence that indexing has a persistent effect on the pricing of securities. According to Masse *et al.* (2000), the information content of inclusion does not fully explain the share price response. Numerous studies examined the effect of changes in the composition of market benchmarks on stock betas (Hau, 2011), the liquidity provision of index funds (Green and Jame, 2011), the demand for component stocks (Kappou *et al.*, 2010), and the home currency (Hau *et al.*, 2010). On the day of index inclusion, NYSE equities saw less severe price effects than NASDAQ stocks, demonstrating the superiority of the expert system over the dealer system in minimizing price effects (Elliott and Warr, 2003).

Cluster 8 and 9: contribution of passive investments to market volatility

This cluster focuses on the *impact exerted by passive investments on market volatility*, which is fueled by both misguided investor decisions and ETF market action driven by arbitrage. This cluster is comprised of 17 articles dating from 2002 to 2018. Herding behavior has been extensively studied during periods of market stress and volatility, and momentum flows during these periods are significant for securities pricing (Goetzmann and Massa, 2002). Investor behavior, on the other hand, is not symmetrical between bull and bear markets (Gleason *et al.*, 2004; Gopane *et al.*, 2023). Stocks with a higher proportion of ETF ownership exhibit much more volatility, which adds another layer of non-diversifiable risk to stock prices (Ben-David *et al.*, 2017). Increased ETF ownership is accompanied by increased trading costs, diminished information acquisition benefits, and an increase in voluntary disclosures (Israeli *et al.*, 2017; Schoenfeld, 2017; Lin *et al.*, 2018). ETFs, hedge funds and other institutional investors provide aggressive liquidity at competitive prices (Aitken *et al.*, 2007). Index funds have been implicated in causing the asset price bubbles and ensuing financial crisis. With the growing popularity of passive investing, the situation is likely to deteriorate further due to increased financialization via ETFs and high-speed automated algorithmic trading, as well as accumulating unfunded obligations (Sornette and Cauwels, 2014). However, the magnitude of the crash is highly dependent on the proportion of uneducated investors and the amount of unsophisticated passive investing in the market (Barlevy and Veronesi, 2003). The effect of stock inclusion in ETFs on price correlation and stock betas remains largely unknown in the EME environment, despite rising cross-sectional commonality in equities. This also highlights the importance of developing innovative risk management solutions.

Nineth cluster is the smallest cluster, consisting of three publications that make significant contributions to the literature on the dynamic role of passive investments in global financial markets. With the growing popularity, passive investments account for a sizable portion of the total AUM in the developed markets. These funds are ready to amass massive positions as their combined ownership in the most widely followed broad indices such as the S&P 500 and DJIA continues to grow at a breakneck pace (Bebchuk and Hirst, 2019). Three mega-funds in the United States control more than 40% of voting rights in companies listed on S&P and dominate the broader index. Their full potential is anticipated to manifest itself in the shift to sustainability, with passive management eclipsing active fund management in socially responsible investing (Chen and Scholtens, 2018; Fisch *et al.*, 2019). The extent to which passive funds contribute to ESG objectives remains an unanswered question.

4.2 Conceptual analysis

The conceptual model illustrated in Figure 6 summarizes a thorough content study of the literature in nine bibliographic clusters. The structure largely illustrates the various antecedents of passive investments and their impact on two major performance attributes, with liquidity, degree of financialization and capital flows serving as mediating factors. Bi-directional arrows suggest potential endogeneity in financial variables, while the link denotes the nature of the relationship studied in the available literature.

The investors' rationale for choosing active funds despite their historical underperformance in comparison to passive benchmarks and index funds has been constantly scrutinized (Goetzmann and Massa, 2003; Lichtenstein *et al.*, 1999; Fortin and Michelson, 2005; Boldin and Cici, 2010). Passive funds are intended to replicate a market-linked benchmark, and deviations from the benchmark define tracking efficiency (Johnson, 2009; Chu, 2011; Rompotis, 2011). Transitions in the macroeconomy dictate market regimes. Furthermore, market volatility is substantially influenced by corporate actions, including investor disclosures and expanding ICT penetration. Expansionary developments tend to generate more pronounced surges, as forecasting errors escalate (Lunde and Zebedee, 2009; Nguyen *et al.*, 2007; Aggarwal *et al.*, 2014). The market for passive investments reflects the cumulative impact of these forces in three broad dimensions: Market liquidity, Capital Flows and Degree of Financialization. Although the term "financialization" began to appear with increasing frequency in the early 1990s, its origins remain elusive. However, since the late 1960s, the fundamental issue of a gravitational movement toward finance in capitalism as a whole has existed. Financialization essentially implies the increasing role of financial motives, financial markets, financial actors, and financial

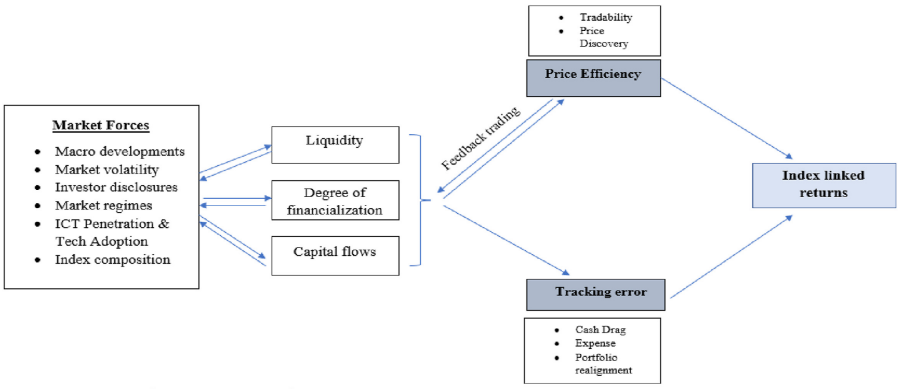


Figure 6.
This figure represents the conceptual structure developed based on analysis of the content of passive investing literature

Source(s): Authors' own work

institutions in the operation of domestic and international economies. This equally holds for passive investments.

ETFs, the most popular passive alternative, allow for secondary market trading. Authorized participants' in-kind redemption and creation processes result in price discovery in ETF markets, defining the pricing efficiency (Shanmughama and Zabiulla, 2012; Petajisto, 2017). The market's liquidity (Ackert and Tian, 2008), as well as overall volatility, determine the success of in-kind creation and redemption processes, thereby influencing the degree of financialization (Irwin and Sanders, 2011, 2012; Michał Falkowski, 2011; Ivanov, 2013). With the financial markets landscape evolving, contemporary issues such as the impact of ICT and social change on ETF performance (Lechman and Marszk, 2015), international diversification (Huang and Lin, 2011) and emerging economies perspectives (Blitz *et al.*, 2012) have become research focal points. Passive investments are gaining popularity and are establishing a dominant position as a source of investible surplus. ETFs have been instrumental in reducing market volatility (Lin and Chiang, 2005) and increasing liquidity (Hegde and McDermott, 2004; Ben-David *et al.*, 2017), and are now the primary source of price discovery (Ivanov, 2013), resulting in long-term convergence of stock betas (Sullivan and Xiong, 2012). Madura and Ngo (2008) concluded that ETFs had a sizable and beneficial effect on the valuation of their constituent equities. Petajisto (2011) coined the term "index turnover cost" to examine the effect of ETF price deviations on investment behavior. Charteris *et al.* (2014) concluded that ETF discounts/premiums encourage feedback trading, whereas Da and Shive (2018) concluded that correlations between ETF returns and stock-lagged beta provide evidence for price reversal. This also reflects the endogeneity of variables such as liquidity, capital flows and financialization degree.

5. Limitations

While the study contributes significantly to the field of knowledge, it has several limitations that must be considered when interpreting its findings and implications. With our emphasis on academic journals, the study analyzed only peer-reviewed journal articles, excluding conference papers, reports, and technical articles. As a result, contributions published in other formats that had an effect on the body of knowledge are excluded. Additionally, our keyword-based approach to article selection may introduce subjectivity bias into the process. While we are confident that our approach resulted in a comprehensive and representative database, our reliance on Elsevier Scopus and Web of Science may have resulted in us overlooking relevant work that is only accessible through other databases. Additionally, specific bibliometric properties may not be time-stable, and certain common distribution patterns of the passive investing literature may still be developing. Furthermore, it is worth noting that the graphs and citation networks were interpreted subjectively. With passive investing's growing popularity and the number of publications at an all-time high, it's clear that intellectual debate in this field has not yet reached saturation. As a result, the study's findings may not be exhaustive, and there is considerable room for additional research, as detailed in Sections 6 and 7.

6. Research implications

With this study, it has been possible to observe and chart the high growth trajectory of passive investing research globally, especially post-US Subprime crisis. The phenomenal growth story of passive investments may be attributable to increasing investor dissatisfaction with active investing which comes at a higher price. The findings of our study have key implications for various stakeholders: investors, regulators and managers, these are discussed in detail below:

6.1 Implication for investors

Despite the widespread adoption of passive investing as an investment strategy, it is not a one-size-fits-all proposition. Market conditions change constantly, and it frequently requires an informed eye to determine when and how much to shift away from active investments and toward passive ones. Active managers may outperform more frequently than passive managers in volatile markets or when the economy is weakening. On the other hand, passive strategies may be preferable when the index's constituent securities move in lockstep or when stocks have more consistent valuations. Depending on the opportunity in different sectors of the capital markets, investors may gain from a combination of passive and active techniques.

While active investing may be more profitable in cases where international diversification is required, such as investing in emerging market stocks, passive investing may make more sense in the domestic investing landscape. Additionally, an active fund manager may be better suited to manage a portfolio of illiquid securities or one that is tailored toward a specific objective, such as limiting the downside risk. It's also worth noting that investors, on average, place a higher premium on risk, return and liquidity than on cost. This is especially true for those with high net worth who have access to private equity, hedge funds and other curated alternative investments. While investing in ETFs, investors should bear in mind the effect of market regimes on TE, as deviations from the benchmark typically increase more during bearish than bullish regimes. Currency ETFs enable investors to implement a carry trade strategy in their portfolios; however, as a word of caution, currency stability and liquidity can play a significant role in international ETFs. Similarly, LETFs may be better suited for dynamic strategies and offer less value to a long-term investor. Lastly, the importance of investor education cannot be underestimated in the name of the highly diversified portfolio when using passive alternatives, for which necessary efforts are required by regulators and investors alike.

6.2 Managerial implications

The inexorable trend to passive investing creates numerous issues for fund management, including fee and revenue pressure, which forces traditional managers to seek new revenue streams, such as illiquid and private assets, which also implies increased portfolio risk. Additionally, the increased transparency and efficiency associated with the ETF market indicates that managers must rethink the entire value chain, beginning with technology and the way investments interact. ETFs outperform closed-ended funds in terms of tracking ability. Both, however, exaggerate their reactions to domestic news. Index funds may be more suitable for international diversification.

According to the research, ETF discounts and premiums promote feedback trading, which may result in panic redemptions during times of market stress. Additionally, managers must develop novel risk management strategies to reduce their reliance on cash holdings, which are a major cause of portfolio TEs. This is truer than before with the more integrated character of the world economy necessitating quicker portfolio realignments as indices reel under pressure. ETFs are now at the forefront of price discovery, and managers must adopt better volatility metrics that take macroeconomic and investment disclosure factors into account. Algorithmic approaches and bootstrap aggregation (Bagging) with superior predicting abilities may prove to be more effective in this area. With increasing sustainability and environmental concerns; however, ETFs are set to overtake active funds in the fulfillment of ESG objectives.

6.3 Implications for regulators

The passive investing phenomenon has shaken the financial industry by assuming dominance on the most closely followed global indices, which contradicts the tenet of

diversification. Passive funds are now having an impact on the creation and composition of the indices they were designed to follow, as trackers' holdings exercise significant influence over voting rights and flow momentum. Passive investments have triggered changes in market structure that are still not fully understood or factored in. Active management and a range of valuation opinions on whether a price is "too low" or "too high" provide much-needed depth to a market as it attempts to strike a delicate balance between demand and supply forces, ensuring liquidity at all price points.

Passive investments, on the other hand, have a straightforward mandate of buying on fund inflows and selling on fund outflows. If passive investments were to dominate, markets would be shallow, with liquidity available only to the extent that prices remained constant. Beyond the current price, the presence of bids and offers rapidly diminishes, increasing the risk of stocks being locked limited up/down and liquidity evaporating as price changes increase in magnitude. As previously stated, NYSE equities saw fewer price repercussions than NASDAQ equities during the index inclusion process, demonstrating the superiority of expert systems over dealer systems based on machine learning. Thus, with passive investing assuming a large role in the financial landscape, regulators will need to devise novel monitoring and discrepancy detection approaches in addition to reinstating a focus on investor education.

7. Conclusion and future scope

The study's objective is to synthesize available knowledge about passive investing and to highlight innovative advancements in this field on a worldwide scale. We analyzed 943 research publications (published between 1998 and 2021) and provided a thorough overview of the characteristics that determined the expansion of passive investment as a result of the twin benefits of diversity and low cost. The collection of studies has major managerial implications, particularly in terms of portfolio allocation, trading methods, price discovery and subsequent investment behavior. As detailed in the cluster analysis section, passive investment has become a dominant force in global financial markets, with passive funds together controlling more than two-fifths of the voting rights in the world's most widely tracked equity indices. ETFs, the most recent financial innovation in passive investing, are currently at the forefront of price discovery. With the increasing popularity of digital platforms, high-frequency trading and market interconnectedness; the impact of market regimes, microstructures and macroeconomic developments on underlying indices and stocks may define fund managers' portfolio optimization strategy. Managers must be alert and creative in weighing the costs and benefits of various options to maximize the performance of their portfolios. Furthermore, this research summarizes the available literature in terms of publication date, citations, geographic scope, intellectual environment, theories applied and research procedures. Additionally, this paper discusses some of the research issues associated with passive investing and suggests future research opportunities in this emerging field of study. Future studies pertaining to the clusters under consideration may be guided by the findings gathered in this study. The effect of contemporary trading platforms and algorithmic trading on price volatility, price formation and liquidity in ETF markets is a relatively uncharted territory within the broader category of market structures. The impact of behavioral characteristics on mispricing, future prices, risk premiums and market efficiency are all aspects that deserve attention within the pricing and market efficiency cluster. Although the number of empirically tested and generalizable studies is on the rise, further research in this area is required. Given the increasing significance of developing nations in international financial markets, there are potential areas for further academic investigation within the portfolio strategies cluster, including the impact of stock inclusions in ETFs, price correlations, stock betas and the variable composition of the indexes

tracked by these funds. Given the heightened awareness surrounding climate risk and sustainability, any discourse concerning ETFs as an alternative investment strategy must also consider the emergence of ETFs with an ESG focus. This represents a substantial aspect that merits further research, especially within the EME context.

Notes

1. The Malaysian government executed an initial monetary and fiscal stimulus package totaling RM 7 billion (USD 0.98 billion). This was succeeded by a more comprehensive stimulus package of RM 15 billion (USD 2.1 billion), which invested in locally-oriented initiatives with multiplier effects and accounted for nearly 9% of the country's GDP. In addition, tax incentives, private finance initiatives, and off-budget ventures were incorporated into the package. The stimulus package was the most comprehensive in the annals of the nation's economy and ranked second among ASEAN nations, behind only Singapore. The circumstances in Malaysia exhibited an unusual pattern, given that the worldwide financial crisis predominantly materialized as an export crisis as opposed to a currency or banking crisis. The significant repercussions were predominantly experienced inside the stock markets, resulting in a rapid outflow of portfolio investments. As a result of avoiding exposure to hazardous property asset bubbles, Malaysian banks encountered limited consequences in other financial sector domains and served as an illustrative case study of insights gained from prior crises.

References

Aber, J.W., Li, D. and Can, L. (2009), "Price volatility and tracking ability of ETFs", *Journal of Asset Management*, Vol. 10 No. 4, pp. 210-221, doi: 10.1057/jam.2009.13.

Acedo, F.J., Barroso, C. and Galan, J.L. (2006), "The resource-based theory: dissemination and main trends", *Strategic Management Journal*, Vol. 27 No. 7, pp. 621-636, doi: 10.1002/smj.532.

Ackert, L.F. and Tian, Y.S. (2008), "Arbitrage, liquidity, and the valuation of exchange traded funds", *Financial Markets, Institutions and Instruments*, Vol. 17 No. 5, pp. 331-362, doi: 10.1111/j.1468-0416.2008.00144.x.

Agapova, A. (2011), "Conventional mutual index funds versus exchange-traded funds", *Journal of Financial Markets*, Vol. 14 No. 2, pp. 323-343, doi: 10.1016/j.finmar.2010.10.005.

Aggarwal, R., Lucey, B.M. and O'Connor, F.A. (2014), "Rationality in precious metals forward markets: evidence of behavioural deviations in the gold markets", *Journal of Multinational Financial Management*, Vols 25-26, pp. 110-130, doi: 10.1016/j.mulfin.2014.06.001.

Ahmed, P. and Nanda, S. (2005), "Performance of enhanced index and quantitative equity funds", *The Financial Review*, Vol. 40 No. 4, pp. 459-479, doi: 10.1111/j.1540-6288.2005.00119.x.

Aitken, M., Almeida, N., de Harris, F.B. and McInish, T.H. (2007), "Liquidity supply in electronic markets", *Journal of Financial Markets*, Vol. 10 No. 2, pp. 144-168, doi: 10.1016/j.finmar.2006.12.001.

Andreu, L., Swinkels, L. and Tjong-A-Tjoe, L. (2013), "Can exchange-traded funds be used to exploit industry and country momentum?", *Financial Markets and Portfolio Management*, Vol. 27 No. 2, pp. 127-148, doi: 10.1007/s11408-013-0207-8.

Avellaneda, M. and Lee, J.-H. (2010), "Statistical arbitrage in the US equities market", *Quantitative Finance*, Vol. 10 No. 7, pp. 761-782, doi: 10.1080/14697680903124632.

Avellaneda, M. and Zhang, S. (2010), "Path-dependence of leveraged ETF returns", *SIAM Journal on Financial Mathematics*, Vol. 1 No. 1, pp. 586-603, doi: 10.1137/090760805.

Ayres, I. and Curtis, Q. (2014), "Beyond diversification: the pervasive problem of excessive fees and 'dominated funds' in 401(k) plans", *Yale LJ*, Vol. 124, p. 1476.

Baks, K.P., Metrick, A. and Wachter, J. (2001), "Should investors avoid all actively managed mutual funds? A study in Bayesian performance evaluation", *The Journal of Finance*, Vol. 56 No. 1, pp. 45-48, doi: 10.1111/0022-1082.00319.

Baltussen, G., Van Bekkum, S. and Da, Z. (2019), "Indexing and stock market serial dependence around the world", *Journal of Financial Economics*, Vol. 132 No. 1, pp. 26-48, doi: 10.1016/j.jfineco.2018.07.016.

Barlevy, G. and Veronesi, P. (2003), "Rational panics and stock market crashes", *Journal of Economic Theory*, Vol. 110 No. 2, pp. 234-263, doi: 10.1016/s0022-0531(03)00039-5.

Bastian, M., Heymann, S. and Jacomy, M. (2009), "Gephi: an open source software for exploring and manipulating networks", *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 3, pp. 361-362, doi: 10.1609/icwsm.v3i1.13937.

Beaulieu, M., Ebrahim, S.K. and Morgan, I.G. (2003), "Does tick size influence price discovery? Evidence from the Toronto Stock Exchange", *Journal of Futures Markets*, Vol. 23 No. 1, pp. 49-66, doi: 10.1002/fut.10053.

Bebchuk, L.A. and Hirst, S. (2019), "The specter of the giant three", *Boston University Law Review*, Vol. 99, pp. 721-741.

Ben-David, I., Franzoni, F. and Moussawi, R. (2017), "Exchange-traded funds", *Annual Review of Financial Economics*, Vol. 9 No. 1, pp. 169-189, doi: 10.1146/annurev-financial-110716-032538.

Ben-David, I., Franzoni, F. and Moussawi, R. (2018), "Do ETFs increase volatility?", *The Journal of Finance*, Vol. 73 No. 6, pp. 2471-2535, doi: 10.1111/jofi.12727.

Beneish, M.D. and Gardner, J.C. (1995), "Information costs and liquidity effects from changes in the Dow Jones industrial average list", *The Journal of Financial and Quantitative Analysis*, Vol. 30 No. 1, p. 135, doi: 10.2307/2331257.

Bhattacharya, U., Loos, B., Meyer, S. and Hackethal, A. (2017), "Abusing ETFs*", *Review of Finance*, Vol. 21 No. 3, pp. 1217-1250, doi: 10.1093/rof/rfw041.

Białkowski, J., Bohl, M.T., Kaufmann, P. and Wisniewski, T.P. (2013), "Do mutual fund managers exploit the Ramadan anomaly? Evidence from Turkey", *Emerging Markets Review*, Vol. 15, pp. 211-232, doi: 10.1016/j.ememar.2013.02.003.

Biktimirov, E.N. (2004), "The effect of demand on stock prices: evidence from index fund rebalancing", *Financial Review*, Vol. 39 No. 3, pp. 455-472, doi: 10.1111/j.0732-8516.2004.00084.x.

Blanco-Mesa, F., Merigó, J.M. and Gil-Lafuente, A.M. (2017), "Fuzzy decision making: a bibliometric-based review", *Journal of Intelligent and Fuzzy Systems*, Vol. 32 No. 3, pp. 2033-2050, doi: 10.3233/jifs-161640.

Blitz, D., Huij, J. and Swinkels, L. (2012), "The performance of European index funds and exchange-traded funds", *European Financial Management*, Vol. 18 No. 4, pp. 649-662, doi: 10.1111/j.1468-036x.2010.00550.x.

Boehmer, B. and Boehmer, E. (2003), "Trading your neighbor's ETFs: competition or fragmentation?", *Journal of Banking and Finance*, Vol. 27 No. 9, pp. 1667-1703, doi: 10.1016/s0378-4266(03)00095-5.

Bogle, J.C. (2014), "The arithmetic of 'all-in' investment expenses", *Financial Analysts Journal*, Vol. 70 No. 1, pp. 13-22, doi: 10.2469/faj.v70.n1.1.

Boldin, M. and Cici, G. (2010), "The index fund rationality paradox", *Journal of Banking and Finance*, Vol. 34 No. 1, pp. 33-43, doi: 10.1016/j.jbankfin.2009.06.018.

Bollen, N.P.B. and Busse, J.A. (2006), "Tick size and institutional trading costs: evidence from mutual funds", *Journal of Financial and Quantitative Analysis*, Vol. 41 No. 4, pp. 915-937, doi: 10.1017/s0022109000002696.

Broman, M.S. (2016), "Liquidity, style investing and excess comovement of exchange-traded fund returns", *Journal of Financial Markets*, Vol. 30, pp. 27-53, doi: 10.1016/j.finmar.2016.05.002.

Brown, J.R., Liang, N. and Weisbenner, S. (2007), "Individual account investment options and portfolio choice: behavioral lessons from 401(k) plans", *Journal of Public Economics*, Vol. 91 No. 10, pp. 1992-2013, doi: 10.3386/w13169.

Brownlees, C.T., Cipollini, F. and Gallo, G.M. (2011), "Intra-daily volume modeling and prediction for algorithmic trading", *Journal of Financial Econometrics*, Vol. 9 No. 3, pp. 489-518, doi: 10.1093/jjfinec/nbq024.

Buetow, G.W. and Henderson, B.J. (2012), "An empirical analysis of exchange-traded funds", *The Journal of Portfolio Management*, Vol. 38 No. 4, pp. 112-127, doi: 10.3905/jpm.2012.38.4.112.

Büyükşahin, B. and Robe, M.A. (2014), "Speculators, commodities and cross-market linkages", *Journal of International Money and Finance*, Vol. 42, pp. 38-70, doi: 10.1016/j.jimonfin.2013.08.004.

Callon, M., Courtial, J.-P., Turner, W.A. and Bauin, S. (1983), "From translations to problematic networks: an introduction to co-word analysis", *Social Science Information*, Vol. 22 No. 2, pp. 191-235, doi: 10.1177/053901883022002003.

Camilleri, S.J., Vassallo, S. and Bai, Y. (2020), "Predictability in securities price formation: differences between developed and emerging markets", *Journal of Capital Markets Studies*, Vol. 4 No. 2, pp. 145-166, doi: 10.1108/jcms-07-2020-0025.

Capelle-Blancard, G. and Coulibaly, D. (2011), "Index trading and agricultural commodity prices: a panel granger causality analysis", *International Economics*, Vols 126-127, pp. 51-71, doi: 10.1016/s2110-7017(13)60036-0.

Charteris, A., Chau, F., Gavriilidis, K. and Kallinterakis, V. (2014), "Premiums, discounts and feedback trading: evidence from emerging markets' ETFs", *International Review of Financial Analysis*, Vol. 35, pp. 80-89, doi: 10.1016/j.irfa.2014.07.010.

Charupat, N. and Miu, P. (2011), "The pricing and performance of leveraged exchange-traded funds", *Journal of Banking and Finance*, Vol. 35 No. 4, pp. 966-977, doi: 10.1016/j.jbankfin.2010.09.012.

Chau, F., Deesomsak, R. and Lau, M.C.K. (2011), "Investor sentiment and feedback trading: evidence from the exchange-traded fund markets", *International Review of Financial Analysis*, Vol. 20 No. 5, pp. 292-305, doi: 10.1016/j.irfa.2011.06.006.

Chen, Y. and Gau, Y. (2009), "Tick sizes and relative rates of price discovery in stock, futures, and options markets: evidence from the Taiwan stock exchange", *Journal of Futures Markets*, Vol. 29 No. 1, pp. 74-93, doi: 10.1002/fut.20319.

Chen, J.-H. and Huang, C.-Y. (2010), "An analysis of the spillover effects of exchange-traded funds", *Applied Economics*, Vol. 42 No. 9, pp. 1155-1168, doi: 10.1080/00036840701721182.

Chen, X. and Scholtens, B. (2018), "The urge to act: a comparison of active and passive socially responsible investment funds in the United States", *Corporate Social Responsibility and Environmental Management*, Vol. 25 No. 6, pp. 1154-1173, doi: 10.1002/csr.1529.

Chen, Y.-L. and Tsai, W.-C. (2017), "Determinants of price discovery in the VIX futures market", *Journal of Empirical Finance*, Vol. 43, pp. 59-73, doi: 10.1016/j.jempfin.2017.05.002.

Chen, H., Noronha, G. and Singal, V. (2006), "Index changes and losses to index fund investors", *Financial Analysts Journal*, Vol. 62 No. 4, pp. 31-47, doi: 10.2469/faj.v62.n4.4185.

Chen, W.-P., Chung, H. and Lien, D. (2016), "Price discovery in the S&P 500 index derivatives markets", *International Review of Economics and Finance*, Vol. 45, pp. 438-452, doi: 10.1016/j.iref.2016.07.008.

Choi, J.J., Laibson, D. and Madrian, B.C. (2010), "Why does the law of one price fail? An experiment on index mutual funds", *The Review of Financial Studies*, Vol. 23 No. 4, pp. 1405-1432, doi: 10.1093/rfs/hhp097.

Chou, R.K. and Chung, H. (2006), "Decimalization, trading costs, and information transmission between ETFs and index futures", *Journal of Futures Markets*, Vol. 26 No. 2, pp. 131-151, doi: 10.1002/fut.20189.

Chu, P.K.-K. (2011), "Study on the tracking errors and their determinants: evidence from Hong Kong exchange traded funds", *Applied Financial Economics*, Vol. 21 No. 5, pp. 309-315, doi: 10.1080/09603107.2010.530215.

Clark, G.L. and Hebb, T. (2004), "Pension fund corporate engagement: the fifth stage of capitalism", *Relations Industrielles*, Vol. 59 No. 1, pp. 142-171, doi: 10.7202/009130ar.

Clifford, C.P., Fulkerson, J.A. and Jordan, B.D. (2014), "What drives ETF flows?", *Financial Review*, Vol. 49 No. 3, pp. 619-642, doi: 10.1111/fire.12049.

Constantinides, G.M., Czerwonko, M., Carsten Jackwerth, J. and Perrakis, S. (2011), "Are options on index futures profitable for risk-averse investors? Empirical evidence", *The Journal of Finance*, Vol. 66 No. 4, pp. 1407-1437, doi: 10.1111/j.1540-6261.2011.01665.x.

Cremers, M., Ferreira, M.A., Matos, P. and Starks, L. (2016), "Indexing and active fund management: international evidence", *Journal of Financial Economics*, Vol. 120 No. 3, pp. 539-560, doi: 10.1016/j.jfineco.2016.02.008.

Cuthbertson, K., Nitzsche, D. and O'Sullivan, N. (2010), "Mutual fund performance: measurement and evidence [1]", *Financial Markets, Institutions and Instruments*, Vol. 19 No. 2, pp. 95-187, doi: 10.1111/j.1468-0416.2010.00156.x.

Da, Z. and Shive, S. (2018), "Exchange-traded funds and asset return correlations", *European Financial Management*, Vol. 24 No. 1, pp. 136-168, doi: 10.1111/eufm.12137.

Daigler, R.T., Hibbert, A.M. and Pavlova, I. (2014), "Examining the return–volatility relation for foreign exchange: evidence from the Euro *VIX*", *Journal of Futures Markets*, Vol. 34 No. 1, pp. 74-92, doi: 10.1002/fut.21582.

Das, S., Kadapakkam, P.-R. and Tse, Y. (2013), "Is carry-trade a viable alternative asset class?", *Journal of International Financial Markets, Institutions and Money*, Vol. 24, pp. 247-257, doi: 10.1016/j.intfin.2012.12.004.

Dedi, L. and Yavas, B.F. (2016), "Return and volatility spillovers in equity markets: an investigation using various GARCH methodologies", *Cogent Economics and Finance*, Vol. 4 No. 1, 1266788, doi: 10.1080/23322039.2016.1266788.

Delcoure, N. and Zhong, M. (2007), "On the premiums of iShares", *Journal of Empirical Finance*, Vol. 14 No. 2, pp. 168-195, doi: 10.1016/j.jempfin.2005.12.004.

Ding, Y. and Cronin, B. (2011), "Popular and/or prestigious? Measures of scholarly esteem", *Information Processing and Management*, Vol. 47 No. 1, pp. 80-96, doi: 10.1016/j.ipm.2010.01.002.

Ding, Y. (2011), "Applying weighted PageRank to author citation networks", *Journal of the American Society for Information Science and Technology*, Vol. 62 No. 2, pp. 236-245, doi: 10.1002/asi.21452.

Donthu, N., Kumar, S. and Pattnaik, D. (2020), "Forty-five years of journal of business research: a bibliometric analysis", *Journal of Business Research*, Vol. 109, pp. 1-14, doi: 10.1016/j.jbusres.2019.10.039.

Ellegaard, O. and Wallin, J.A. (2015), "The bibliometric analysis of scholarly production: how great is the impact?", *Scientometrics*, Vol. 105 No. 3, pp. 1809-1831, doi: 10.1007/s11192-015-1645-z.

Elliott, W.B. and Warr, R.S. (2003), "Price pressure on the NYSE and Nasdaq: evidence from S&P 500 index changes", *Financial Management*, Vol. 32 No. 3, p. 85, doi: 10.2307/3666385.

Elton, E.J., Gruber, M.J. and Blake, C.R. (1996), "The persistence of risk-adjusted mutual fund performance", *The Journal of Business*, Vol. 69 No. 2, p. 133, doi: 10.1086/209685.

Elton, E.J., Gruber, M.J., Comer, G. and Li, K. (2002), "Spiders: where are the Bugs?", *The Journal of Business*, Vol. 75 No. 3, pp. 453-472, doi: 10.1086/339891.

Elton, E.J., Gruber, M.J. and Busse, J.A. (2004), "Are investors rational? Choices among index funds", *The Journal of Finance*, Vol. 59 No. 1, pp. 261-288, doi: 10.1111/j.1540-6261.2004.00633.x.

Elton, E.J., Gruber, M.J. and Blake, C.R. (2007), "Participant reaction and the performance of funds offered by 401(k) plans", *Journal of Financial Intermediation*, Vol. 16 No. 2, pp. 249-271, doi: 10.1016/j.jfi.2006.12.002.

Elton, E.J., Gruber, M.J. and Blake, C.R. (2012), "Does mutual fund size matter? The relationship between size and performance", *Review of Asset Pricing Studies*, Vol. 2 No. 1, pp. 31-55, doi: 10.1093/rapstu/ras001.

Engle, R.F. and Sarkar, D. (2006), "Premiums-discounts and exchange traded funds", *The Journal of Derivatives*, Vol. 13 No. 4, pp. 27-45, doi: 10.3905/jod.2006.635418.

Estrada, J. (2008), "Fundamental indexation and international diversification", *The Journal of Portfolio Management*, Vol. 34 No. 3, pp. 93-109, doi: 10.3905/jpm.2008.706247.

Fahling, E.J., Steurer, E. and Sauer, S. (2019), "Active vs Passive funds—an empirical analysis of the German equity market", *Journal of Financial Risk Management*, Vol. 08 No. 02, pp. 73-91, doi: 10.4236/jfrm.2019.82006.

Fan, J., Furger, A. and Xiu, D. (2016), "Incorporating global industrial classification standard into portfolio allocation: a simple factor-based large covariance matrix estimator with high-frequency data", *Journal of Business and Economic Statistics*, Vol. 34 No. 4, pp. 489-503, doi: 10.1080/07350015.2015.1052458.

Fernholz, R., Garvy, R. and Hannon, J. (1998), "Diversity-weighted indexing", *The Journal of Portfolio Management*, Vol. 24 No. 2, pp. 74-82, doi: 10.3905/jpm.24.2.74.

Fisch, J., Hamdani, A. and Solomon, S.D. (2019), "The new titans of wall street: a theoretical framework for passive investors", *University of Pennsylvania Law Review*, Vol. 168 No. 1, pp. 17-72.

Fong, W.M., Wong, W.K. and Lean, H.H. (2005), "International momentum strategies: a stochastic dominance approach", *Journal of Financial Markets*, Vol. 8 No. 1, pp. 89-109, doi: 10.1016/j.finmar.2004.08.001.

Fortin, R. and Michelson, S. (2005), "Active international mutual fund management; can managers neat the index?", *Managerial Finance*, Vol. 31 No. 1, pp. 41-51, doi: 10.1108/03074350510769451.

Frino, A. and Gallagher, D.R. (2001), "Tracking S&P 500 index funds", *The Journal of Portfolio Management*, Vol. 28 No. 1, pp. 44-55, doi: 10.3905/jpm.2001.319822.

Frino, A., Gallagher, D.R., Neubert, A.S. and Oetomo, T.N. (2004), "Index design and implications for index tracking", *The Journal of Portfolio Management*, Vol. 30 No. 2, pp. 89-95, doi: 10.3905/jpm.2004.319934.

Gao, L., Han, Y., Zhengzi Li, S. and Zhou, G. (2018), "Market intraday momentum", *Journal of Financial Economics*, Vol. 129 No. 2, pp. 394-414, doi: 10.1016/j.jfineco.2018.05.009.

Gasbarro, D., Wong, W.-K. and Kenton Zumwalt, J. (2007), "Stochastic dominance analysis of iShares", *The European Journal of Finance*, Vol. 13 No. 1, pp. 89-101, doi: 10.1080/13518470601025243.

Gastineau, G.L. (2004), "The benchmark index ETF performance problem", *The Journal of Portfolio Management*, Vol. 30 No. 2, pp. 96-103, doi: 10.3905/jpm.2004.319935.

Gil-Bazo, J. and Ruiz-Verdú, P. (2009), "The relation between price and performance in the mutual fund industry", *The Journal of Finance*, Vol. 64 No. 5, pp. 2153-2183, doi: 10.1111/j.1540-6261.2009.01497.x.

Gilbert, C.L. and Pfuderer, S. (2014), "The role of index trading in price formation in the grains and oilseeds markets", *Journal of Agricultural Economics*, Vol. 65 No. 2, pp. 303-322.

Gleason, K.C., Mathur, I. and Peterson, M.A. (2004), "Analysis of intraday herding behavior among the sector ETFs", *Journal of Empirical Finance*, Vol. 11 No. 5, pp. 681-694, doi: 10.1016/j.jempfin.2003.06.003.

Global Asset Management (2021), available at: bcg.com

Goetzmann, W.N. and Massa, M. (2002), "Daily momentum and contrarian behavior of index fund investors", *The Journal of Financial and Quantitative Analysis*, Vol. 37 No. 3, p. 375, doi: 10.2307/3594985.

Goetzmann, W.N. and Massa, M. (2003), "Index funds and stock market growth*", *The Journal of Business*, Vol. 76 No. 1, pp. 1-28, doi: 10.1086/344111.

Golec, J. (2003), "Regulation and the rise in asset-based mutual fund management fees", *Journal of Financial Research*, Vol. 26 No. 1, pp. 19-30, doi: 10.1111/1475-6803.00042.

Gopane, T.J., Moyo, N.T. and Setaka, L.F. (2023), "Emerging market analysis of passive and active investing under bear and bull market conditions", *Journal of Capital Markets Studies*, Vol. ahead-of-print ahead-of-print, doi: 10.1108/jcms-03-2023-0008.

Green, T.C. and Jame, R. (2011), "Strategic trading by index funds and liquidity provision around S&P 500 index additions", *Journal of Financial Markets*, Vol. 14 No. 4, pp. 605-624, doi: 10.1016/j.finmar.2011.02.004.

Gruber, M.J. (1996), "Another Puzzle: the growth in actively managed mutual funds", *The Journal of Finance*, Vol. 51 No. 3, pp. 783-810, doi: 10.1111/j.1540-6261.1996.tb02707.x.

Guercio, D.D. and Reuter, J. (2014), "Mutual fund performance and the incentive to generate alpha", *The Journal of Finance*, Vol. 69 No. 4, pp. 1673-1704, doi: 10.1111/jofi.12048.

Hamilton, J.D. and Wu, J.C. (2015), "Effects of index-fund investing on commodity futures prices", *International Economic Review*, Vol. 56 No. 1, pp. 187-205, doi: 10.1111/iere.12099.

Hansen, P.R., Huang, Z. and Shek, H.H. (2012), "Realized GARCH: a joint model for returns and realized measures of volatility", *Journal of Applied Econometrics*, Vol. 27 No. 6, pp. 877-906, doi: 10.1002/jae.1234.

Harper, J.T., Madura, J. and Schnusenberg, O. (2006), "Performance comparison between exchange-traded funds and closed-end country funds", *Journal of International Financial Markets, Institutions and Money*, Vol. 16 No. 2, pp. 104-122, doi: 10.1016/j.intfin.2004.12.006.

Hasbrouck, J. (2003), "Intraday price formation in U.S. equity index markets", *The Journal of Finance*, Vol. 58 No. 6, pp. 2375-2400, doi: 10.1046/j.1540-6261.2003.00609.x.

Haslam, J., Ji, J. and Sun, H. (2018), "Towards a well-functioning stock market in context", *Journal of Capital Markets Studies*, Vol. 2 No. 1, pp. 21-36, doi: 10.1108/jcms-01-2018-0003.

Hau, H. (2011), "Global versus local asset pricing: a new test of market integration", *Review of Financial Studies*, Vol. 24 No. 12, pp. 3891-3940, doi: 10.1093/rfs/hhr094.

Hau, H., Massa, M. and Peress, J. (2010), "Do demand curves for currencies Slope down? Evidence from the MSCI global index change", *Review of Financial Studies*, Vol. 23 No. 4, pp. 1681-1717, doi: 10.1093/rfs/hhp095.

Hegde, S.P. and McDermott, J.B. (2004), "The market liquidity of DIAMONDS, Q's, and their underlying stocks", *Journal of Banking and Finance*, Vol. 28 No. 5, pp. 1043-1067, doi: 10.1016/s0378-4266(03)00043-8.

Hendershott, T. and Jones, C.M. (2005), "Island goes dark: transparency, fragmentation, and regulation", *Review of Financial Studies*, Vol. 18 No. 3, pp. 743-793, doi: 10.1093/rfs/hhi013.

Hirsch, J.E. (2005), "An index to quantify an individual's scientific research output", *Proceedings of the National Academy of Sciences*, Vol. 102, pp. 16569-16572, doi: 10.1073/pnas.0507655102.

Hortacsu, A. and Syverson, C. (2004), "Product differentiation, search costs, and competition in the mutual fund industry: a case study of S&P 500 index funds", *The Quarterly Journal of Economics*, Vol. 119 No. 2, pp. 403-456, doi: 10.1162/0033553041382184.

Houge, T. and Wellman, J. (2007), "The use and abuse of mutual fund expenses", *Journal of Business Ethics*, Vol. 70 No. 1, pp. 23-32, doi: 10.1007/s10551-006-9077-6.

Hu, K., Wu, H., Qi, K., Yu, J., Yang, S., Yu, T., Zheng, J. and Liu, B. (2018), "A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model", *Scientometrics*, Vol. 114 No. 3, pp. 1031-1068, doi: 10.1007/s11192-017-2574-9.

Huang, M.-Y. and Lin, J.-B. (2011), "Do ETFs provide effective international diversification?", *Research in International Business and Finance*, Vol. 25 No. 3, pp. 335-344, doi: 10.1016/j.ribaf.2011.03.003.

Hughen, J.C. and Mathew, P.G. (2009), "The efficiency of international information flow: evidence from the ETF and CEF prices", *International Review of Financial Analysis*, Vol. 18 Nos 1-2, pp. 40-49, doi: 10.1016/j.irfa.2008.12.001.

Inci, A.C. (2018), "Cointegration and causality in capital markets", *Journal of Capital Markets Studies*, Vol. 2 No. 1, pp. 82-94, doi: 10.1108/jcms-03-2018-0009.

Ippolito, R.A. (1989), "Efficiency with costly information: a study of mutual fund performance, 1965-1984", *The Quarterly Journal of Economics*, Vol. 104 No. 1, p. 1, doi: 10.2307/2937832.

Irwin, S.H. and Sanders, D.R. (2011), "Index funds, financialization, and commodity futures markets", *Applied Economic Perspectives and Policy*, Vol. 33 No. 1, pp. 1-31, doi: 10.1093/aepp/ppq032.

Irwin, S.H. and Sanders, D.R. (2012), "Testing the masters hypothesis in commodity futures markets", *Energy Economics*, Vol. 34 No. 1, pp. 256-269, doi: 10.1016/j.eneco.2011.10.008.

Israeli, D., Lee, C.M.C. and Sridharan, S.A. (2017), "Is there a dark side to exchange traded funds? An information perspective", *Review of Accounting Studies*, Vol. 22 No. 3, pp. 1048-1083, doi: 10.1007/s11142-017-9400-8.

Ivanov, S.I. (2013), "The influence of ETFs on the price discovery of gold, silver and oil", *Journal of Economics and Finance*, Vol. 37 No. 3, pp. 453-462, doi: 10.1007/s12197-011-9205-8.

Jaeger, L. and Wagner, C. (2005), "Factor modeling and benchmarking of hedge funds: can passive investments in hedge fund strategies deliver?", *The Journal of Alternative Investments*, Vol. 8 No. 3, pp. 9-36, doi: 10.3905/jai.2005.608030.

Jares, T.E. and Lavin, A.M. (2004), "Japan and Hong Kong exchange-traded funds (ETFs): discounts, returns, and trading strategies", *Journal of Financial Services Research*, Vol. 25 No. 1, pp. 57-69, doi: 10.1023/b:fina.0000008665.55707.ab.

Johnson, W.F. (2009), "Tracking errors of exchange traded funds", *Journal of Asset Management*, Vol. 10 No. 4, pp. 253-262, doi: 10.1057/jam.2009.10.

Joshi, G. and Dash, R.K. (2024), "Exchange-traded funds and the future of passive investments: a bibliometric review and future research agenda", *Future Business Journal*, Vol. 10 No. 1, 17, doi: 10.1186/s43093-024-00306-8.

Kappou, K., Brooks, C. and Ward, C. (2010), "The S&P500 index effect reconsidered: evidence from overnight and intraday stock price performance and volume", *Journal of Banking and Finance*, Vol. 34 No. 1, pp. 116-126, doi: 10.1016/j.jbankfin.2009.07.008.

Keim, D.B. (1999), "An analysis of mutual fund design: the case of investing in small-cap stocks", *Journal of Financial Economics*, Vol. 51 No. 2, pp. 173-194, doi: 10.1016/s0304-405x(98)00049-x.

Kessler, M.M. (1963), "Bibliographic coupling between scientific papers", *American Documentation*, Vol. 14 No. 1, pp. 10-25, doi: 10.1002/asi.5090140103.

Khorana, A., Nelling, E. and Trester, J.J. (1998), "The emergence of country index funds", *The Journal of Portfolio Management*, Vol. 24 No. 4, pp. 78-84, doi: 10.3905/jpm.1998.409651.

Kostovetsky, L. (2003), "Index mutual funds and exchange-traded funds", *The Journal of Portfolio Management*, Vol. 29 No. 4, pp. 80-92, doi: 10.3905/jpm.2003.319897.

Krause, T. and Tse, Y. (2013), "Volatility and return spillovers in Canadian and U.S. industry ETFs", *International Review of Economics and Finance*, Vol. 25, pp. 244-259, doi: 10.1016/j.iref.2012.07.009.

Kumar, A., Niessen-Ruenzi, A. and Spalt, O.G. (2015), "What's in a name? Mutual fund flows when managers have foreign-sounding names", *Review of Financial Studies*, Vol. 28 No. 8, pp. 2281-2321, doi: 10.1093/rfs/hhv017.

Lechman, E. and Marszk, A. (2015), "ICT technologies and financial innovations: the case of exchange traded funds in Brazil, Japan, Mexico, South Korea and the United States", *Technological Forecasting and Social Change*, Vol. 99, pp. 355-376, doi: 10.1016/j.techfore.2015.01.006.

Lee, C.-C. and Chen, M.-P. (2020), "Do natural disasters and geopolitical risks matter for cross-border country exchange-traded fund returns?", *The North American Journal of Economics and Finance*, Vol. 51, 101054, doi: 10.1016/j.najef.2019.101054.

Lettau, M. and Madhavan, A. (2018), "Exchange-traded funds 101 for economists", *Journal of Economic Perspectives*, Vol. 32 No. 1, pp. 135-154, doi: 10.1257/jep.32.1.135.

Leung, T. and Ward, B. (2015), "The golden target: analyzing the tracking performance of leveraged gold ETFs", *Studies in Economics and Finance*, Vol. 32 No. 3, pp. 278-297, doi: 10.1108/sef-01-2015-0009.

Levy, A. and Lieberman, O. (2013), "Overreaction of country ETFs to US market returns: intraday vs daily horizons and the role of synchronized trading", *Journal of Banking and Finance*, Vol. 37 No. 5, pp. 1412-1421, doi: 10.1016/j.jbankfin.2012.03.024.

Li, M. and Zhao, X. (2014), "Impact of leveraged ETF trading on the market quality of component stocks", *The North American Journal of Economics and Finance*, Vol. 28, pp. 90-108, doi: 10.1016/j.najef.2014.02.001.

Lichtenstein, D.R., Kaufmann, P.J. and Bhagat, S. (1999), "Why consumers choose managed mutual funds over index funds: hypotheses from consumer behavior", *Journal of Consumer Affairs*, Vol. 33 No. 1, p. 187, doi: 10.1111/j.1745-6606.1999.tb00766.x.

Lin, C.-C. and Chiang, M.-H. (2005), "Volatility effect of ETFs on the constituents of the underlying Taiwan 50 Index", *Applied Financial Economics*, Vol. 15 No. 18, pp. 1315-1322, doi: 10.1080/09603100500389630.

Lin, Y., Mao, Y. and Wang, Z. (2018), "Institutional ownership, peer pressure, and voluntary disclosures", *The Accounting Review*, Vol. 93 No. 4, pp. 283-308, doi: 10.2308/accr-51945.

Lo, A.W. (2007), "Efficient markets hypothesis", in Blume, L. and Durlauf, S. (Eds), *The New Palgrave: A Dictionary of Economics*, 2nd ed., Palgrave Macmillan, available at: https://ssrn.com/abstract=991509

Lunde, A. and Zebedee, A.A. (2009), "Intraday volatility responses to monetary policy events", *Financial Markets and Portfolio Management*, Vol. 23 No. 4, pp. 383-399, doi: 10.1007/s11408-009-0114-1.

Lynch, A.W. and Mendenhall, R.R. (1997), "New evidence on stock price effects associated with changes in the S&P 500 index", *The Journal of Business*, Vol. 70 No. 3, pp. 351-383, doi: 10.1086/209722.

Madhavan, A. (2003), "The Russell reconstitution effect", *Financial Analysts Journal*, Vol. 59 No. 4, pp. 51-64, doi: 10.2469/faj.v59.n4.2545.

Madhavan, A. (2012), "Exchange-traded funds, market structure, and the flash crash", *Financial Analysts Journal*, Vol. 68 No. 4, pp. 20-35, doi: 10.2469/faj.v68.n4.6.

Madura, J. and Ngo, T. (2008), "Impact of ETF inception on the valuation and trading of component stocks", *Applied Financial Economics*, Vol. 18 No. 12, pp. 995-1007, doi: 10.1080/09603100701335424.

Main, S., Irwin, S.H., Sanders, D.R. and Smith, A. (2018), "Financialization and the returns to commodity investments", *Journal of Commodity Markets*, Vol. 10, pp. 22-28, doi: 10.1016/j.jcomm.2018.05.004.

Malkiel, B.G. (2005), "Reflections on the efficient market hypothesis: 30 Years later", *The Financial Review*, Vol. 40 No. 1, pp. 1-9, doi: 10.1111/j.0732-8516.2005.00090.x.

Malkiel, B.G. and Radisich, A. (2001), "The growth of index funds and the pricing of equity securities", *The Journal of Portfolio Management*, Vol. 27 No. 2, pp. 9-21, doi: 10.3905/jpm.2001.319788.

Martyn, J. (1964), "Bibliographic coupling", *Journal of Documentation*, Vol. 20 No. 4, p. 236, doi: 10.1108/eb026352.

Mas-Tur, A., Modak, N.M., Merigó, J.M., Roig-Tierno, N., Geraci, M. and Capecchi, V. (2019), "Half a century of quality & quantity: a bibliometric review", *Quality and Quantity*, Vol. 53 No. 2, pp. 981-1020, doi: 10.1007/s11135-018-0799-1.

Masse, I., Hanrahan, R., Kushner, J. and Martinello, F. (2000), "The effect of additions to or deletions from the TSE 300 Index on Canadian share prices", *Canadian Journal of Economics/Revue Canadienne d'économique*, Vol. 33 No. 2, pp. 341-359, doi: 10.1111/0008-4085.00019.

Meade, N. and Salkin, G.R. (1990), "Developing and maintaining an equity index fund", *Journal of the Operational Research Society*, Vol. 41 No. 7, pp. 599-607, doi: 10.2307/2583437.

Merigó, J.M., Muller, C., Modak, N.M. and Laengle, S. (2019), "Research in production and operations management: a university-based bibliometric analysis", *Global Journal of Flexible Systems Management*, Vol. 20 No. 1, pp. 1-29, doi: 10.1007/s40171-018-0201-0.

Michał Falkowski, M.F. (2011), "Financialization of commodities", *Contemporary Economics*, Vol. 5 No. 4, p. 4, doi: 10.5709/ce.1897-9254.24.

Mikhaylov, A., Sokolinskaya, N. and Lopatin, E. (2019), "Retraction: asset allocation in equity, fixed-income and cryptocurrency on the base of individual risk sentiment", *Investment Management and Financial Innovations*, Vol. 16 No. 2, pp. 171-181, doi: 10.21511/imfi.16(2).2019.15.

Miralles-Quirós, J.L. and Miralles-Quirós, M.M. (2019), "Are alternative energies a real alternative for investors?", *Energy Economics*, Vol. 78, pp. 535-545, doi: 10.1016/j.eneco.2018.12.008.

Moore, D.A., Kurtzberg, T.R., Fox, C.R. and Bazerman, M.H. (1999), "Positive illusions and forecasting errors in mutual fund investment decisions", *Organizational Behavior and Human Decision Processes*, Vol. 79 No. 2, pp. 95-114, doi: 10.1006/obhd.1999.2835.

Nerur, S.P., Rasheed, A.A. and Natarajan, V. (2008), "The intellectual structure of the strategic management field: an author co-citation analysis", *Strategic Management Journal*, Vol. 29 No. 3, pp. 319-336, doi: 10.1002/smj.659.

Nguyen, V., Van Ness, B.F. and Van Ness, R.A. (2007), "Short-and long-term effects of multimarket trading", *Financial Review*, Vol. 42 No. 3, pp. 349-372, doi: 10.1111/j.1540-6288.2007.00175.x.

Nugroho, B.A. (2021), "Dynamic risk-based optimization on cryptocurrencies", *Journal of Capital Markets Studies*, Vol. 5 No. 1, pp. 28-48, doi: 10.1108/jcms-01-2021-0002.

O'Neal, E.S. (2004), "Purchase and redemption patterns of US equity mutual funds", *Financial Management*, pp. 63-90.

Ordu, B.M., Oran, A. and Soytas, U. (2018), "Is food financialized? Yes, but only when liquidity is abundant", *Journal of Banking and Finance*, Vol. 95, pp. 82-96, doi: 10.1016/j.jbankfin.2017.06.001.

Osterhoff, F. and Kaserer, C. (2016), "Determinants of tracking error in German ETFs – the Role of Market Liquidity", *Managerial Finance*, Vol. 42 No. 5, pp. 417-437.

Pattnaik, D., Hassan, M.K., Kumar, S. and Paul, J. (2020), "Trade credit research before and after the global financial crisis of 2008 – a bibliometric overview", *Research in International Business and Finance*, Vol. 54, 101287, doi: 10.1016/j.ribaf.2020.101287.

Pedersen, L.H. (2018), "Sharpening the arithmetic of active management", *Financial Analysts Journal*, Vol. 74 No. 1, pp. 21-36, doi: 10.2469/faj.v74.n1.4.

Petajisto, A. (2011), "The index premium and its hidden cost for index funds", *Journal of Empirical Finance*, Vol. 18 No. 2, pp. 271-288, doi: 10.1016/j.jempfin.2010.10.002.

Petajisto, A. (2017), "Inefficiencies in the pricing of exchange-traded funds", *Financial Analysts Journal*, Vol. 73 No. 1, pp. 24-54, doi: 10.2469/faj.v73.n1.7.

Pézier, J. and White, A. (2008), "The relative merits of alternative investments in passive portfolios", *The Journal of Alternative Investments*, Vol. 10 No. 4, pp. 37-49, doi: 10.3905/jai.2008.705531.

Poterba, J.M. and Shoven, J.B. (2002), "Exchange-traded funds: a new investment option for taxable investors", *American Economic Review*, Vol. 92 No. 2, pp. 422-427.

Pullen, T., Benson, K. and Faff, R. (2014), "A comparative analysis of the investment characteristics of alternative gold assets", *Abacus*, Vol. 50 No. 1, pp. 76-92, doi: 10.1111/abac.12023.

Randhawa, K., Wilden, R. and Hohberger, J. (2016), "A bibliometric review of open innovation: setting a research agenda", *Journal of Product Innovation Management*, Vol. 33 No. 6, pp. 750-772, doi: 10.1111/jpim.12312.

Rompotis, G.G. (2009), "Interfamily competition on index tracking: the case of the vanguard ETFs and index funds", *Journal of Asset Management*, Vol. 10 No. 4, pp. 263-278, doi: 10.1057/jam.2009.11.

Rompotis, G.G. (2011), "Predictable patterns in ETFs' return and tracking error", *Studies in Economics and Finance*, Vol. 28 No. 1, pp. 14-35, doi: 10.1108/10867371111110534.

Sanders, D.R., Irwin, S.H. and Merrin, R.P. (2010), "The adequacy of speculation in agricultural futures markets: too much of a good thing?", *Applied Economic Perspectives and Policy*, Vol. 32 No. 1, pp. 77-94, doi: 10.1093/aepp/ppp006.

Sarkar, S.S., Dutta, S. and Dutta, P. (2013), "A review of Indian index funds", *Global Business Review*, Vol. 14 No. 1, pp. 89-98, doi: 10.1177/0972150912466445.

Schlusche, B. (2009), "Price Formation in spot and futures markets: exchange traded funds vs index futures", *The Journal of Derivatives*, Vol. 17 No. 2, pp. 26-40, doi: 10.3905/jod.2009.17.2.026.

Schoenfeld, J. (2017), "The effect of voluntary disclosure on stock liquidity: new evidence from index funds", *Journal of Accounting and Economics*, Vol. 63 No. 1, pp. 51-74, doi: 10.1016/j.jacceco.2016.10.007.

Scholes, M.S. (1972), "The market for securities: substitution versus price pressure and the effects of information on share prices", *The Journal of Business*, Vol. 45 No. 2, p. 179, doi: 10.1086/295444.

Shanmugham, R. and Zabiulla (2012), "Pricing efficiency of nifty BeES in bullish and bearish markets", *Global Business Review*, Vol. 13 No. 1, pp. 109-121, doi: 10.1177/097215091101300107.

Sharpe, W.F. (1991), "The arithmetic of active management", *Financial Analysts Journal*, Vol. 47 No. 1, pp. 7-9, doi: 10.2469/faj.v47.n1.7.

Sherrill, D.E., Shirley, S.E. and Stark, J.R. (2017), "Actively managed mutual funds holding passive investments: what do ETF positions tell us about mutual fund ability?", *Journal of Banking and Finance*, Vol. 76, pp. 48-64, doi: 10.1016/j.jbankfin.2016.11.025.

Shin, S. and Soydemir, G. (2010), "Exchange-traded funds, persistence in tracking errors and information dissemination", *Journal of Multinational Financial Management*, Vol. 20 Nos 4-5, pp. 214-234, doi: 10.1016/j.mulfin.2010.07.005.

Shleifer, A. (1986), "Do demand curves for stocks Slope down?", *The Journal of Finance*, Vol. 41 No. 3, pp. 579-590, doi: 10.2307/2328486.

Shynkevich, A. (2016), "Predictability in bond returns using technical trading rules", *Journal of Banking and Finance*, Vol. 70, pp. 55-69, doi: 10.1016/j.jbankfin.2016.06.010.

Small, H. (1973), "Co-citation in the scientific literature: a new measure of the relationship between two documents", *Journal of the American Society for Information Science*, Vol. 24 No. 4, pp. 265-269, doi: 10.1002/asi.4630240406.

Sorensen, E.H., Miller, K.L. and Samak, V. (1998), "Allocating between active and passive management", *Financial Analysts Journal*, Vol. 54 No. 5, pp. 18-31, doi: 10.2469/faj.v54.n5.2209.

Sornette, D. and Cauwels, P. (2014), "1980-2008: the illusion of the perpetual money machine and what it bodes for the future", *Risks*, Vol. 2 No. 2, pp. 103-131, doi: 10.3390/risks2020103.

Stoll, H.R. and Whaley, R.E. (1990), "The dynamics of stock index and stock index futures returns", *The Journal of Financial and Quantitative Analysis*, Vol. 25 No. 4, p. 441, doi: 10.2307/2331010.

Stoll, H.R. and Whaley, R.E. (2011), "Commodity index investing: speculation or diversification?", *The Journal of Alternative Investments*, Vol. 14 No. 1, pp. 50-60, doi: 10.3905/jai.2011.14.1.050.

Sullivan, R.N. and Xiong, J.X. (2012), "How index trading increases market vulnerability", *Financial Analysts Journal*, Vol. 68 No. 2, pp. 70-84, doi: 10.2469/faj.v68.n2.7.

Takeda, A., Niranjan, M., Gotoh, J. and Kawahara, Y. (2013), "Simultaneous pursuit of out-of-sample performance and sparsity in index tracking portfolios", *Computational Management Science*, Vol. 10 No. 1, pp. 21-49, doi: 10.1007/s10287-012-0158-y.

Tang, H. and Xu, X.E. (2013), "Solving the return deviation conundrum of leveraged exchange-traded funds", *Journal of Financial and Quantitative Analysis*, Vol. 48 No. 1, pp. 309-342, doi: 10.1017/s0022109012000622.

Tang, H. and Xu, X.E. (2017), "A review of fixed-income indexing and index investing", *The Journal of Wealth Management, Pageant Media US*, Vol. 19 No. 4, pp. 85-103, doi: 10.3905/jwm.2017.19.4.085.

Thompson, J.R., Baggett, L.S., Wojciechowski, W.C. and Williams, E.E. (2006), "Nobels for nonsense", *Journal of Post Keynesian Economics*, Vol. 29 No. 1, pp. 3-18.

Tranfield, D., Denyer, D. and Smart, P. (2003), "Towards a methodology for developing evidence-informed management knowledge by means of systematic review", *British Journal of Management*, Vol. 14 No. 3, pp. 207-222, doi: 10.1111/1467-8551.00375.

Tsay, M.-Y. (2009), "Citation analysis of Ted Nelson's works and his influence on hypertext concept", *Scientometrics*, Vol. 79 No. 3, pp. 451-472, doi: 10.1007/s11192-008-1641-7.

Tse, Y. and Erenburg, G. (2003), "Competition for order flow, market quality, and price discovery in the nasdaq 100 index tracking stock", *Journal of Financial Research*, Vol. 26 No. 3, pp. 301-318, doi: 10.1111/1475-6803.00060.

Tse, Y. and Martinez, V. (2007), "Price discovery and informational efficiency of international iShares funds", *Global Finance Journal*, Vol. 18 No. 1, pp. 1-15, doi: 10.1016/j.gfj.2007.02.001.

Tse, Y., Bandyopadhyay, P. and Shen, Y. (2006), "Intraday price discovery in the DJIA index markets", *Journal of Business Finance and Accounting*, Vol. 33 Nos 9-10, pp. 1572-1585, doi: 10.1111/j.1468-5957.2006.00639.x.

Vespro, C. (2006), "Stock price and volume effects associated with compositional changes in European stock indices", *European Financial Management*, Vol. 12 No. 1, pp. 103-127, doi: 10.1111/j.1354-7798.2006.00312.x.

Wong, K.H.Y. and Shum, W.C. (2010), "Exchange-traded funds in bullish and bearish markets", *Applied Economics Letters*, Vol. 17 No. 16, pp. 1615-1624, doi: 10.1080/13504850903085035.

Woolridge, J.R. and Ghosh, C. (1986), "Institutional trading and security prices: the case of changes in the composition of the S&P 500 index", *Journal of Financial Research*, Vol. 9 No. 1, pp. 13-24, doi: 10.1111/j.1475-6803.1986.tb00432.x.

Wu, L.-C., Chou, S.-C., Yang, C.-C. and Ong, C.-S. (2007), "Enhanced index investing based on goal programming", *The Journal of Portfolio Management*, Vol. 33 No. 3, pp. 49-56, doi: 10.3905/jpm.2007.684753.

Zupic, I. and Čater, T. (2015), "Bibliometric methods in management and organization", *Organizational Research Methods*, Vol. 18 No. 3, pp. 429-472, doi: 10.1177/1094428114562629.

Further reading

Ellis, M.E. (2004), "Purchase and redemption patterns of US equity mutual funds", *CFA Digest*, Vol. 34 No. 4, pp. 66-68, doi: 10.2469/dig.v34.n4.1577.

Friesen, G.C. and Sapp, T.R.A. (2007), "Mutual fund flows and investor returns: an empirical examination of fund investor timing ability", *Journal of Banking and Finance*, Vol. 31 No. 9, pp. 2796-2816, doi: 10.1016/j.jbankfin.2007.01.024.

Ivanov, S.I., Jones, F.J. and Zaima, J.K. (2013), "Analysis of DJIA, S&P 500, S&P 400, NASDAQ 100 and Russell 2000 ETFs and their influence on price discovery", *Global Finance Journal*, Vol. 24 No. 3, pp. 171-187, doi: 10.1016/j.gfj.2013.10.005.

Massey, D.S., Opare, M.A., Wallach, J.D., Ross, J.S. and Krumholz, H.M. (2020), "Assessment of preprint policies of top-ranked clinical journals", *JAMA Network Open*, Vol. 3 No. 7, p. e2011127, doi: 10.1001/jamanetworkopen.2020.11127.

Sanders, D.R. and Irwin, S.H. (2017), "Bubbles, froth and facts: another look at the masters hypothesis in commodity futures markets", *Journal of Agricultural Economics*, Vol. 68 No. 2, pp. 345-365, doi: 10.1111/1477-9552.12191.

Corresponding author
Priya Malhotra can be contacted at: priyamalhotra.17@fms.edu, priyamalhotra@dtu.ac.in

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

Towards Optimal Resource Allocation: A Multi-agent DRL based Task Offloading Approach in Multi-UAV-Assisted MEC Networks

MUHAMMAD NAQQASH TARIQ [1], JINGYU WANG [1], (Senior Member, IEEE), SALMAN RAZA [2], MOHAMMAD SIRAJ [3], (Senior Member, IEEE), MAJID ALTAMIMI [3], (Member, IEEE), and SAIFULLAH MEMON[1].

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications.
[2]Department of Computer Science, National Textile University, Faisalabad 37610, Pakistan.
[3]Department of Electrical Engineering, College of Engineering, King Saud University, Riyadh 11543, Saudi Arabia.

The authors are grateful to the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunication (BUPT), Beijing, China, for providing the Laboratory Services. The authors extend their appreciation to the Deputyship for Research and Innovation, 433 Ministry of Education, Saudi Arabia for funding this research (IFKSUOR3-459).

ABSTRACT The application of UAV-aided MEC well-suited for the execution of the data-intensive and latency-sensitive tasks in the infrastructure-deprived regions. However, the growing number of UAVs and smart devices causing a major difficulty in the devising an effective scheme for the task offloading and resource allocation in multi-UAV-aided MEC networks. Furthermore, the resource deficient environments unable to sustain prolonged resource-intensive activities, additional complexities are posed on the optimum utilization of the resources. In this paper, we introduced a multi-agent deep reinforcement learning scheme for the task offloading in the multi-UAV-assisted networks (MUAVDRL). In this configuration, the mobile users fetch computational resources from the UAVs with the goal of minimizing the computation cost which incorporates both the energy consumption and the computation delay. Initially, we start with the optimization problem which is defined as the minimizing the computational costs. Through modelling it as MDP, we aim to reduce the computational costs for mobile users. Leveraging the dynamic and high-dimensional nature of the challenge, the MUAVDRL algorithm solves this problem efficiently. Comprehensive simulation results exhibit the efficacy and superiority of our projected framework when compared to existing state-of-the-art methods, illustrating its potential in the practice.

INDEX TERMS DRL, MEC, Resource allocation, Task offloading, UAV.

I. INTRODUCTION

The swift progression of the Internet of Things (IoTs) and information and communication technology has propelled smart applications like virtual reality (VR), augmented reality(AR), and autonomous driving [1] into the spotlight. As a result, the demand for minimal latency and considerable computational power has risen with the emergence of these applications, thereby posing substantial challenges for IoT mobile devices. These problems can be solved by Mobile edge computing (MEC) which provides the resources like the cloud at the network edge.. However, traditional MEC servers are costly to deploy and may not be viable in areas lacking infrastructure. To reinforce MEC's adaptability, UAV-assisted MEC has emerged as a promising paradigm [2]. However, prevalent solutions for optimization problems lean on centralized frameworks, reliant on the computational capacity of a central controller. With rising IoT and UAV numbers, complexity, and computational costs surging, they are urging the need for distributed decision-making methods to alleviate this computational burden in UAV-assisted MEC networks.

While mobile devices within UAV-assisted networks increasingly run real-time applications, their operational scope remains significantly restricted due to limited computational resources [3]. Introducing task-offloading strategies has expanded their capabilities by enabling application execution task delegation to nearby computational nodes [4]. However, this enhancement faces challenges, notably the potential hin-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022

drance caused by insufficient mobile battery energy [5]. This limitation often leads to service disruptions, as mobile applications may prematurely terminate [6]. The uninterrupted function of contemporary wireless networks necessitates prolonged battery life for sustained energy availability, alongside effective energy utilization by edge devices [7]. These issues collectively impede the network's ability to process real-time applications within designated time slots. Moreover, [8] focuses on optimizing caching in a cellular network where mobile devices are powered by harvested energy. It aims to minimize the proportion of file segments retrieved from the base station (BS) by optimizing file placement on devices. In another study, Lu et al. [9] investigate the placement of unmanned aerial vehicle base stations (UAV-BSs) in mobile UAV networks to minimize UAV-recall-frequency (UAV-RF) and enhance energy efficiency. Considering various power consumptions and ground user density, they employ a pattern formation system to capture the unstable nature of user density. Additionally, [10] delves into trajectory planning for UAVs in UAV-aided IoT networks with mobile sensor nodes (SNs), focusing on priority-oriented time-sensitive data collection. By assigning different delay sensitivities to each SN, the aim is to minimize UAV energy consumption and SN average delay through trajectory optimization. A proposed heuristic trajectory planning algorithm leverages RL but may introduce oscillations in convergence models, potentially impacting stability and effectiveness in specific scenarios.

Deep Reinforcement Learning (DRL) has gained significant traction, especially within network resource allocation problem domains. Its success in video games [11], continuous action control [12], and now in network resource allocations is noteworthy. Researchers are increasingly exploring DRL's potential in these domains, recognizing that optimal resource allocation is essentially an optimal decision-making challenge. When network dynamics exhibit certain regularities, DRL agents can discern these patterns and learn effective policies accordingly [13]. Chen et al. [14] delved into designing computation offloading policies in Edge Computing (EC) through deep Q-networks, integrating task queue state, energy queue state, and channel state information. Meanwhile, Ye et al. [15] focused on resource allocation a DRL-based scheme within V2V communications, treating V2V links as agents making informed decisions regarding optimal power levels and sub-band for transmissions. Sun et al. [16] examined the computation migration as a crucial issue in vehicular cloud scenario, employing polynomial estimation for value functions and SARSA algorithm training to address missions with linear inter-dependency topology. Conversely, Zhu et al. [17] explored research on UAV-supported edge computing for missions with different inter-dependence topology structures using only single agent Actor-Critic (AC) algorithm for selecting target UAVs, albeit without accounting for parallel transmission and energy constraints. Additionally, Min et al. [18] came up with another offloading algorithm based on RL for IoT devices with energy harvesting, considering factors like current battery levels, previous radio transmission rates, and predicted harvested energy. However, the authors overlooked the time delay factor concerning energy consumption, potentially limiting the comprehensive nature of their approach.

There has been substantial research into UAVs aiding mobile edge computing, covering aspects like energy efficiency [19], trajectory design [20], and the joint optimization of communication and computation [21]. However, few of these studies account for scenarios involving multiple UAVs. A single UAV, constrained by a limited payload, possesses restricted computing resources. Moreover, several recent studies have explored multi-UAV-assisted MEC. specifically, Huang et al. [22] delve into offloading scenarios within multi-UAV-enabled aerial edge computing. The authors addressed a scenario where users have interdependent tasks, and multiple UAVs process these tasks centrally. Using a multi-objective optimization approach, they aim to optimize completion time and energy balance among UAVs. However, this study overlooked computing resource allocation, and practical constraints, like limited resources, could indirectly affect its optimization objectives.Consequently, UAV-assisted mobile edge computing necessitates a collective effort among UAVs to share the computation workload. Diverse computing capabilities among UAVs result in varying processing and communication times. Additionally, differences in energy conditions between UAVs and mobile users directly impact offloading outcomes. Within the dynamic network environment, efficiently offloading tasks for mobile users during ground network emergencies poses a significant challenge. The local ground MEC server might become overwhelmed as mobile users demand varied resources within the UAV network. Moreover, the mobile nature of some users causes them to fall outside the network coverage. Considering UAV-assisted networks' complex and dynamic conditions and resources, the task offloading policy should be carefully designed to minimize computation costs.

Due to the limitations of preceding works and the challenges stated above, in this article, we introduced a multi-agent DRL into a UAV-enabled network to propose intelligent task offloading. We offer task offloading and resource allocation, where mobile users offload their tasks to UAVs to perform task offloading. We develop the problem and transform it into a Markov decision process (MDP) model to attain optimum offloading decision policies. Given the presence of multi-objective problems and the dynamic nature inherent in UAV-enabled network environments, we incorporate the capabilities of advanced DRL techniques. This approach allows us to effectively grasp the dynamic network conditions, paving the way to devise an optimal task offloading algorithm that efficiently manages interactions between mobile users and UAVs.

This work's pivotal contributions can be encapsulated as follows:

1) We proposed a MUAVDRL scheme for task offloading among mobile devices within a multi-UAV-assisted

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022

network. The primary aim is to minimize computation costs, specifically targeting energy consumption and computation delay. This scheme empowers mobile devices to acquire environmental insights, enabling them to empower optimal task-offloading decisions within the network.

2) We developed a resource allocation and task offloading problem for mobile devices into a Markov MDP model and utilize a multi-agent model-free deep deterministic policy gradient algorithm for multi-UAV-assisted networks (MUAVDDPG). This method manages the expansive and continuous action space to maximize long-term rewards and determine the most efficient task-offloading strategy.
3) We extensively simulate and compare the performance of our proposed algorithm against three benchmark algorithms. These simulations demonstrate that our proposed algorithm significantly minimizes computation costs within a dynamic network setting.

The rest of this paper is arranged as follows. In Section II, the system model is presented. In Section III, the problem is formulated is presented. The MUAVDRL scheme and the algorithm are presented in Section IV. The simulation results are given in Section V. Finally, Section VI presents the conclusion.

II. SYSTEM MODEL

TABLE 1: List of Notations

Notations	Description
M	Set of UAVs
N	Set of mobile users
D_n	Set of computation task demand
S_n	Task Size
C_n	Required CPU cycles
τ_n	Delay Constraint
W	Bandwidth
d_{mn}	Distance between UAV m and user n
G_{mn}	Channel gain from user n to UAV m
r_{mn}	Transmission rate between mobile user n and UAV m
$T^{off}_{n,m}$	Transmission delay of user n offloading tasks to UAV m
$E^{off}_{n,m}$	Energy consumption of user n for offloading to UAV m
$T^{re}_{n^{(i)},m}$	Queuing Delay
$T^{tot}_{n,m}$	Total Delay
C^{tot}	Total system cost

The illustration in Fig. 1 delves into a comprehensive analysis of a Multi-UAV-assisted MEC network, constituting a collective of M UAVs and N mobile users. The UAVs are concisely labeled within this network structure in the set $M = \{1, 2, \ldots, m, \ldots, M\}$. At the same time, the mobile users are categorised explicitly in $N = \{1, 2, \ldots, n, \ldots, N\}$, which provides a structured and scalable representation, allowing for systematic coordination and efficient communication among the various entities present in the network. Owing to limited computational capabilities, mobile devices must transfer their computation tasks to UAVs for processing. Employing a comprehensive offloading model is essential to minimize the time and energy consumption of the user's equipment. Each UAV has a maximum capacity to connect with up to N_{max} mobile users concurrently, while each user is limited to offloading tasks to only one UAV during a specific time.

Each mobile user $n \in N$ contributes a computational task described by $D_n = (S_n, C_n, \tau_n)$, here S_n, C_n, and τ_n represent the task size, necessry for CPU-cycle frequencies, and for the imposed delay constraint, respectively. The parameter $\alpha_n \in \{1, 2, \ldots, M\}$ serves as the association policy for user n, indicating the selection of a UAV by user n for offloading purposes. This association policy is crucial in optimizing resource allocation, enabling each user to select a UAV strategically based on factors such as proximity, resource availability, and computation efficiency, ensuring efficient task offloading and timely processing within the multi-UAV environment. The notation used in the system model are illustrated in Table 1.

A. CHANNEL MODEL

The overall system bandwidth, denoted as W, is distributed among M unmanned aerial vehicles (UAVs), each utilizing a bandwidth of W_m. To prevent interference among these UAVs, the combined bandwidth across all UAVs must not surpass the overall system bandwidth, i.e., $\sum_{m=1}^{M} W_m \leq W$. Moreover, an equal distribution of bandwidth allocation is assumed among users connected to a particular UAV. This means that users connected with UAV m evenly divide the overall bandwidth W_m allocated to that specific UAV m. To ensure optimal utilization of bandwidth, the collective bandwidth used by all UAVs should not fall below the total system bandwidth W [23], which can be represented as follows:

$$\sum_{m \in \mathcal{M}} W_m = W, 0 < W_m < W. \tag{1}$$

The coordinates defining the position of UAV $m \in M$ and user $n \in N$ are specified as $L_m = (x_m, y_m, h)$ and $L_n = (x_n, y_n, 0)$, respectively. To determine the Euclidean distance in a three-dimensional space between two points UAV m and user n, an approach detailed in [2] is adopted, which can be expressed as:

$$d_{mn} = \sqrt{(x_m - x_n)^2 + (y_m - y_n)^2 + h^2} \tag{2}$$

Under the assumption, communication link among UAVs and mobile users is primarily Line of Sight (LoS), as highlighted in [24], the channel gain between mobile user n and UAV m is defined as per [25]:

$$G_{mn} = (d_{mn})^{-\lambda}, \tag{3}$$

where, the symbol λ illustrates the path loss exponent, serving as a parameter in the characterization of path loss for the communication link between the mobile user and the UAV.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022

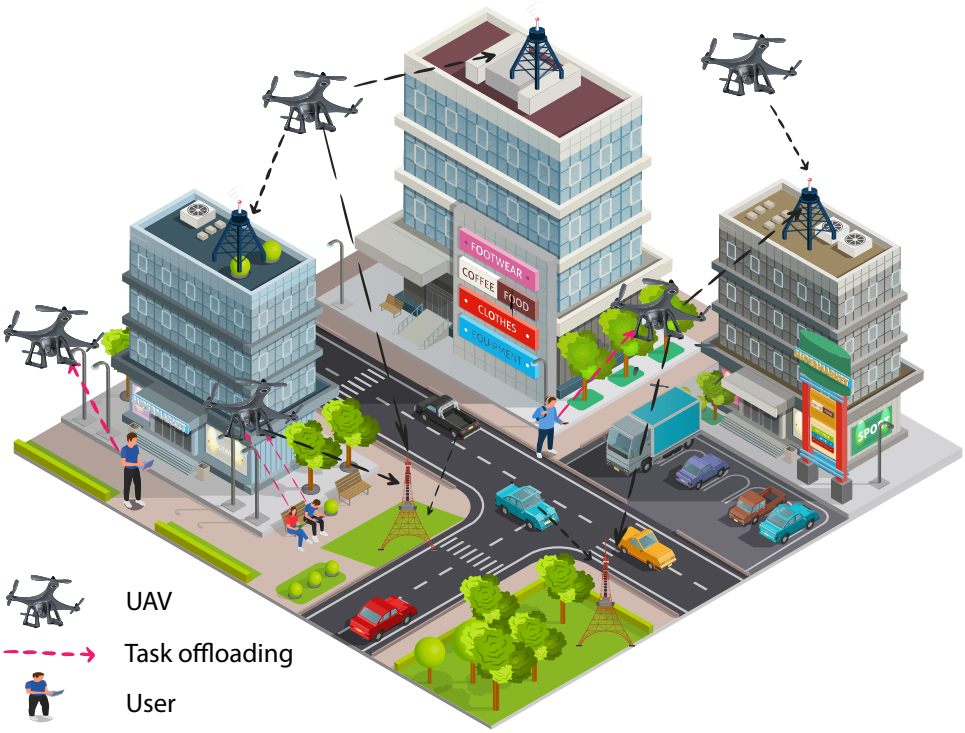


FIGURE 1: System model showcasing Multi-UAV-assisted MEC network.

B. OFFLOADING MODEL

The expression for the transmission rate between UAV m and mobile user n for the offloaded task D_n to UAV m is represented as:

$$r_{mn} = \frac{W_m}{|\mathcal{N}_m|}\log_2\left(1 + \frac{p_n G_{mn}}{N_0}\right), \tag{4}$$

where, W_m stands for the total bandwidth allocated to UAV m. $\mathcal{N}_m$ signifies the count of users associated to UAV m, with N_m representing this set of users as $N_m = n1, n2, \ldots, n|N_m|$. ensures an equitable distribution of available bandwidth resources among the users connected to UAV m. Additionally, the variable p_n refers to the transmission power of user n, while N_0 encompasses the background noise within the system.

The time taken for user n to transfer tasks to UAV m, referred to as the transmission delay, can be represented as:

$$T^{off}_{n,m} = \frac{S_n}{r_{mn}}. \tag{5}$$

The energy consumption related to user n during the process of offloading tasks to its respective UAV can be formulated as:

$$E^{off}_{n,m} = p_n T^{off}_n. \tag{6}$$

C. QUEUING MODEL

Once all computation tasks are received from their corresponding users, each UAV m organizes these tasks based on their order of arrival within the set of user $\mathcal{N}_m = \{n^1, n^2, \ldots, n^{|\mathcal{N}_m|}\}$. Consequently, the computation delay for user n at MEC m includes both queuing and computation delay. The queuing delay, crucially dependent on the order of task arrival, aggregates the computational delays of all tasks handled prior to the current task in the sequence. Tasks arriving earlier experience a longer queuing delay as they wait for previously arrived tasks to be processed. For instance, considering user $n^{(i)} \in \mathcal{N}_m$, processing delay pertaining to $n^{(i)}$ is outlined in [26].

$$T^{re}_{n^{(i)},m} = \begin{cases} \frac{C_{n^{(i)}}}{f_m}, i = 1 \\ T^{re}_{n^{(i-1)},m} + \frac{C_{n^{(i)}}}{f_m}, i > 1 \end{cases}, \tag{7}$$

Therefore, accounting for the available computation frequency of UAV m, denoted as f_m, the comprehensive latency incurred in processing tasks from user n encompasses both the queuing delay and the computational time. As defined in (7) the remaining processing time $T^{re}_{n^{(i)},m}$ for tasks from user $n^{(i)}$ at UAV m. It comprises two cases: firstly, when $i = 1$, indicating the first task from user n, the remaining processing time is straightforwardly computed as the computational time of the initial task $C_{n^{(i)}}$ divided by the computation frequency of UAV m as f_m. Subsequently, for tasks beyond the first $i > 1$, the remaining processing time is determined by adding the computational time of the current task $C_{n^{(i)}}$ divided by f_m to the remaining processing time of the previous task $T^{re}_{n^{(i-1)},m}$. This formulation enables tracking of the time required for task completion, considering

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022

both the computational demands of individual tasks and the processing capabilities of the UAV. Hence, the total latency is calculated as the sum of the queuing delay and the time required for computation on UAV m, which is defined as follows.

$$T_{n,m}^{tot} = T_{n,m}^{off} + T_{n,m}^{re}, \tag{8}$$

The variable $\beta_n = \{0, 1\}$ functions as a binary indicator, signifying whether user n gratifies the stipulated delay requirement. When $T_{n,m}^{tot} \leq \tau_n \cdot \beta_n = 1$, it indicates that the user meets the specified delay constraints. This condition ensures that the total time taken for task processing by user n at UAV m does not exceed the designated maximum time threshold, thus adhering to the latency requirements.

Consequently, the total amount of the system cost amalgamates the energy consumption and time utilization from all users whose computational tasks are completed successfully. The representation of the total amount of the system cost is formulated as the summation of energy and time costs across all users in the system whose task computations are within the stipulated time constraints:

$$\begin{aligned} C^{tot} &= \delta^e E^{tot} + \delta^t T^{tot} \\ &= \delta^e \sum_{n\in\mathcal{N}} \sum_{m\in\mathcal{M}} \beta_n p_n T_{n,m}^{tot} + \delta^t \sum_{n\in\mathcal{N}} \sum_{m\in\mathcal{M}} \beta_n T_{n,m}^{tot}. \end{aligned} \tag{9}$$

III. PROBLEM FORMULATION

The primary objective is to minimize the cost of the system in multi-UAV-assisted network scenarios. The system cost is expressed as energy consumption and computation delay. For this, we formulate an optimization problem to minimize the system cost of the system through optimizing task offloading strategy and resource allocation to eventually reduce the overall system cost and is mathematically expressed as:

$$\begin{aligned} &P1: \min_{T,E} C^{tot} \\ &s.t. \sum_{m\in\mathcal{M}} W_m = W, \forall m, && \text{(10a)} \\ &\alpha_n \in \{1, 2, \ldots, m, \ldots, M\}, \forall n, m, && \text{(10b)} \\ &|\mathcal{N}_m| \leq N_{\max}, , \forall m, && \text{(10c)} \\ &p_n \in \{1, 2, \ldots, n, \ldots, N\}, \forall n. && \text{(10d)} \\ &E_n^{off} \leq E_n^{max}, \quad \forall n. && \text{(10e)} \\ &T_n^{tot} \leq T_n^{max}, \quad \forall n && \text{(10f)} \end{aligned}$$

The optimization problem (*P1*) revolves around key variables: the user association vector, transmit power vector, and UAV bandwidth allocation vector. These variables are defined as follows:

1) User Association Vector (α): The $\alpha = [\alpha_1, , \alpha_n, \ldots, \alpha_N]$ represents the user association vector. Each α_n denotes the choice of a UAV by user n for task offloading.
2) Transmit Power Vector (p): The $p = [p_1, \ldots, p_n, \ldots, p_N]$ represents the transmit power vector. The p_n corresponds to the transmission power of individual users.
3) UAV Bandwidth Allocation Vector (W): The $W = [W_1, \ldots, W_m, \ldots, W_M]$ denotes the UAV bandwidth allocation vector. W_m stands for the total bandwidth allocated to each specific UAV m in the system.

These variables set the foundation for the optimization problem, which is further defined by the following constraints:

1) Bandwidth Restriction ($\sum_{m\in\mathcal{M}} W_m = W, \forall m$): Constraint (10a) ensures that the aggregate bandwidth usage across all UAVs does not exceed the total system bandwidth W.
2) User Association ($\alpha_n \in \{1, 2, \ldots, m, \ldots, M\}, \forall n, m$): Constraint (10b) represents the user association indicator governed by the user association vector α. It imposes a limit on the maximum number of users a UAV can associate with, denoted as $N_{\max}$.
3) Maximum User Association ($|\mathcal{N}_m| \leq N_{\max}, , \forall m$): Constraint (10c) regulates the number of users connected to each UAV, aiding in workload management and optimization for individual UAVs within the network.
4) Transmit Power Range ($p_n \in \{1, 2, \ldots, n, \ldots, N\}, \forall n$): Constraint (10d) signifies the permissible range for users' transmit power.
5) Energy Limitation ($E_n^{off} \leq E_n^{max}, \quad \forall n$): Constraint (10e) outlines mobile users' energy limitation, ensuring that each computing device's battery energy remains adequate.
6) Time Constraint ($T_n^{tot} \leq T_n^{max}, \quad \forall n$): Constraint (10f) sets the time constraint for mobile users, stipulating that the total time T_n^{tot} should not surpass the maximum delay constraint for the task, denoted as T_n^{max}.

The proliferation of mobile devices is experiencing rapid exponential growth, coupled with the dynamic network scalability, thereby elevating the complexity of problem P1. This complexity stems from its classification as a mixed-integer and non-convex optimization problem. It revolves around the offloading decision vector α and features a nonconvex objective function, a characteristic that renders it NP-hard and poses challenges in directly deriving an optimal solution. An algorithm in polynomial time would not be able to unveil the best decision.Hence, we are providing a simplified DRL algorithm that minimizes the system offloading cost while maximizing resource allocation efficiency.

IV. MULTI-AGENT DRL SCHEME

A. MUAVDRL BASED TASK OFFLOADING

In this paper, we explore the task offloading paradigm within a multi-UAV-supported network operating within a dynamic environment where multiple agents are actively involved. The complexities of optimizing solutions in such a setting using traditional single-agent approaches have been acknowledged

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022

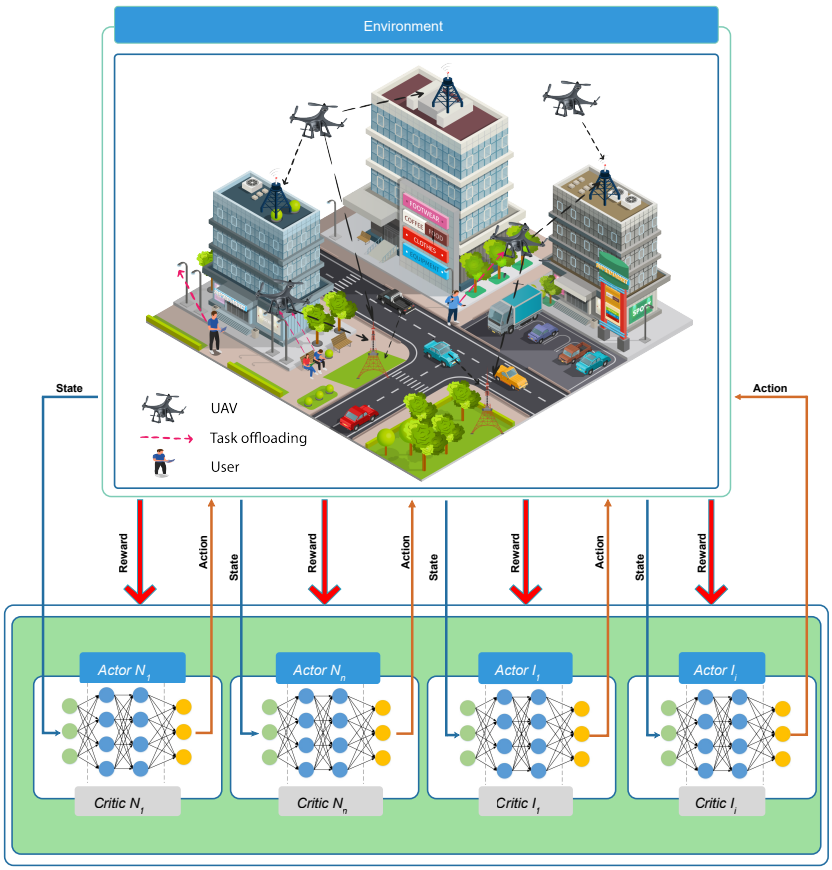


FIGURE 2: MUAVDRL Scheme for multi-UAV-assisted MEC Network.

as challenging [27]. To overcome these hurdles, our contribution introduces the MUAVDRL algorithm, tailored explicitly to address multiple problems within a multi-agent system. This algorithm is a strategic response to the complexities of optimizing task offloading across a network supported by multiple UAV agents.

We have formulated the problems within a Markov decision process framework in a multi-agent setting. Our approach introduces a MUAVDRL-based task offloading scheme aimed at minimizing the computation costs of mobile users while concurrently maximizing the rewards accrued by UAVs. Within this scheme, mobile devices and UAVs function as agents, engaging in a competitive interplay to accomplish their objectives. As illustrated in Fig. 2, agents compete during each stage, gaining experiential insights from others by observing the environmental state. The rewards acquired by UAVs and mobile users constitute the system's overall reward. To execute the MUAVDRL method effectively, we define the state space, action space, and rewards as follows:

- **State space:** The learning agents within the system acquire experiential knowledge and refine their decision-making policies by observing the environmental states. Their decisions are selected by optimization objectives within the environment, relying on the state space $s(t)$. This state space $s(t) = (r_{mn}, E^{off}_{n,m}, f_m, p_n)$, comprising elements such as the transmission rate r_{mn} between user n and UAV m, user energy consumption $E^{off}_{n,m}$ during task offloading, UAV computation frequency f_m, and user transmit power p_n, serves as the cornerstone for these agents' adaptive decision-making, enabling them to optimize strategies based on the current environmental conditions.

- **Action space:** The vector of offloading decision $B = [\beta_1, \beta_2, \ldots, \beta_N]$, vector of transmit power $p = [p_1, \ldots, p_n, \ldots, p_N]$, and computation resource allocation is represented as $F = [f_1, f_2, \ldots, f_N]$. The complete action vector at time 't' is indicated as $a(t) = [\beta_1, p_1, f_1, \ldots, \beta_N, p_N, f_N]$.

- **Reward function:** The overarching aim of the reward function is to enhance offloading decisions, mitigating computation costs encompassing delays and energy consumption. This optimization is crucial for ensuring QoS while extending the battery lifespan of mobile users. The reward function for mobile users is formulated as:

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022

$$r_{n,m}(t) = -(\delta^e E^{tot} + \delta^t T^{tot}). \quad (10)$$

The reward function for UAVs is defined as:

$$r_m(t) = R_m(t). \quad (11)$$

The comprehensive reward function for the system is formulated as:

$$r(t) = \sum_{n\in\mathcal{N}} \sum_{m\in\mathcal{M}} \Big(r_{n,m}(t) + r_m(t)\Big). \quad (12)$$

B. MUAVDRL ALGORITHM

Fig. 2 illustrates the MUAVDRL framework within a multi-UAV-supported network. The MUAVDDPG algorithm, an adapted actor-critic network incorporating the DQN method, demonstrates its efficacy in managing dynamic environments. Particularly suited for cooperative policy learning among multiple agents operating within a continuous action space, this algorithm stands as a valuable tool. It is essential to highlight that this system facilitates centralized training of the action-value function, known as the critic network while promoting decentralized execution. The critic function utilizes action policies from other agents, enabling each agent to decide its actions autonomously based on individual strategies and observations. The update of policy parameters is carried out individually by each agent within the framework.

The scheme involves a collection of J agents, encompassing both UAVs denoted as M_m and mobile users denoted as N_n, each equipped with distinctive observations, actions, and reward mechanisms. Within this context, each agent, indexed as J among the J agents, possesses a distinct set of states denoted as $\mathcal{S} = \{S_1, S_2, \ldots, S_J\}$, an array of actions denoted as $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_J\}$, and a series of observations represents as $\mathcal{O} = \{\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_J\}$. In this scheme, agents are empowered to select an action utilizing a stochastic policy $\pi_{(\theta_j)} : \mathcal{O}_j \times \mathcal{A}_j \rightarrow [0, 1]$ to select an action based on its own observations and available actions. The transition to the subsequent state is facilitated by the state transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A}_j \times, \ldots, \times \mathcal{A}_J \rightarrow S$. Every agent receives rewards centered on their current states and the actions taken, depicted as $r_j : \mathcal{S} \times \mathcal{A}_j \rightarrow \mathcal{R}$. Additionally, each agent privately acquires their specific observation $\mathcal{O}_j : \mathcal{S} \rightarrow \mathcal{O}_j$. Each agent j operates with the primary goal of maximizing its expected cumulative rewards, denoted as $\mathbb{E}\left\{\sum_{j=0}^{\infty} \gamma^j r_j, t + k\right\}$. Here, $r_j, t + k$ signifies the reward obtained by agent j at a time k episodes into the future, considering the discount factor γ. Within the MUAVDDPG framework, the fundamental concept revolves around learning a centralized critic function, namely the action-value function, designated as $Q_j^{\pi}(\mathcal{J}, \alpha_1, \ldots, \alpha_J)$ for the j-th agent. This function involves representations of actions from all agents $\{\alpha_1, \ldots, \alpha_J\} \in \mathcal{A}$, and each agent possesses comprehensive information regarding global training weights.

Moreover, the collective policies' set utilized by agents is denoted as $\pi = \pi_1, \ldots, \pi_J$, each accompanied by its respective policy parameter represented as $\theta = \theta_1, \ldots, \theta_J$. This centralized learning approach underpins the MUAVDDPG framework, facilitating coordinated learning among multiple agents within the system. Within the MUAVDDPG algorithm, updates are performed on both the actor and critic-network upon completing every training episode. More precisely, the actor-network undergoes an update using a gradient method computed in the following manner:

$$\nabla_{\theta_j}(\theta_j) = \mathbb{E}_{\mathcal{J},a\sim\mathcal{D}}[\nabla_{a_j}\theta_j^{\mu}(\mathcal{J}, a_1, \ldots, a_J) \nabla_{\theta_j}\mu\theta_j(\mathcal{O}_j)|_{aj} = \mu\theta_j(\mathcal{O}_j)], \quad (13)$$

where, the variable $\mathcal{D}$ signifies the replay buffer, responsible for storing tuples of experiences encountered by all agents, consisting of $(\mathcal{J}, \mathcal{J}', a_1, \ldots, a_J, r_1, \ldots, r_J)$. The update process for the critic function θ_j^{μ} is represented as follows:

$$\mathcal{L}(\theta_j) = \mathbb{E}_{\mathcal{J},\mathcal{J}',a_1,\ldots,a_J,r_1,\ldots,r_J}[(\theta_j^{\mu}(\mathcal{J}, a_1, \ldots, a_J) - y)^2], \quad (14)$$

where, $y = r_j + \gamma\theta j^{\mu'}(\mathcal{J}', a1', \ldots, a_J')|ak' = \mu k'(\mathcal{O}k)$ denotes the computation for updating the critic function. Here, the target policies' set, characterized by delayed parameters, is represented as θ_j', represented as θ_j', is depicted as $\mu' = \mu\theta_1', \ldots, \mu\theta_J'$. This calculation involves incorporating the rewards received by agent j alongside evaluating the next state-action values utilizing the target policies, considering the delayed parameters.

Algorithm 1 outlines the proposed MUAVDDPG algorithm tailored for task offloading within a multi-UAV-supported network. It comprises two essential procedures: Collecting the observed data and the training process. The algorithm commences with an initialization phase (lines 1-4), establishing the replay buffer and initializing parameters for the actor-network and critic-network along with their respective weights. This phase also sets the groundwork by defining the total number of episodes and training steps for the agents (lines 5-10). This preliminary step ensures that the algorithm starts with a defined structure and clear parameters before the actual training begins. Collection of observed data by agents: Execution of actions, reward acquisition, and generation of new states (lines 11-21). The experiences gained are stored within the experience replay buffer. Training phase (lines 22-28): Policy training involves batch sampling from the replay buffer. Subsequently, the actor-network and critic-network undergo updates based on randomly selected samples. This algorithm orchestrates the collection of data from the environment by agents, storing experiences for later training. During the training phase, samples from these experiences to update the actor and critic networks, facilitating policy learning within the multi-agent system.

C. COMPUTATIONAL COMPLEXITY

The proposed MUAVDDPG algorithm outlines the computational complexities for actor and critic networks. For the

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022

Algorithm 1 MUAVDDPG Algorithm for Task Offloading

Initialization:
Replay buffer $\mathcal{D}_j^{loc}$ at UAV and User.
Initialization:
Actor and critic networks parameters are initialized with θ
for episode $= 1$ to 3000 **do**
  Initialize the states $S = \{s_1, s_2, \ldots, s_J\}$
  for $t = 1$ to 200 **do**
    Estimate the resource requirements of the user $(E_{n,m}^{off}, f_n)$
    Estimate the available resource of UAV (f_m)
    Each agent receives an initial state. $s(t) = (r_{mn}, E_{n,m}^{off}, f_m, p_n)$
    Every agent j makes a stochastic decision a_j according to the policy π_{θ_j}
given the state s_j with probability ε
    Agents perform the action $a(t) = \{a_1(t), a_2(t), \ldots, a_J(t)\}$, encompassing $a = [\beta_1, p_1, f_1, \ldots, \beta_N, p_N, f_N]$
    The rewards observed are denoted as $r(t) = \{r_1(t), r_2(t), \ldots, r_J(t)\}$
    The new state $s_j(t+1)$ is denoted by s'_j
    store the tuples $\{s_j(t), a_j(t), r_j(t), s'_j\}$ in $\mathcal{D}_j^{loc}$
    $s_J \leftarrow s'_j$
    for agent $j = 1$ to J **do**
      Mini-batch of H samples tuples (s^k, a^k, r^k, s'^k) from D_j^{loc}
      Set $y^k = r_j^k + \gamma Q_j^{\pi'}\left(S'^k, a'_1, \ldots, a'_J\right)|_{a'_j = \pi'_j(s_j^k)}$
      Update the critic-network as (14)
      Update actor-network as (13)
    end for
    Update the target network's parameters for each agent j
    $\theta'_j \leftarrow \tau\theta_j + (1-\tau)\theta'_j$
  end for
end for

actor-network with P layers and the critic network with U layers, the computational complexity of each layer is expressed as $\mathcal{O}\left(Z_{p-1}^a Z_p^a + Z_p^a Z_{p+1}^a\right)$ and $\mathcal{O}\left(Z_{u-1}^c Z_u^c + Z_u^c Z_{u+1}^c\right)$ respectively. Consequently, the overall training complexity of the proposed algorithm can be expressed as:

$$\mathcal{O}\left(J * T * E * \left(\sum_{p=2}^{P-1}\left(Z_{g-1}^a Z_p^a + Z_p^a Z_{p+1}^a\right) + \sum_{u=2}^{U-1}\left(Z_{u-1}^c Z_u^c + Z_u^c Z_{u+1}^c\right)\right)\right) \quad (15)$$

Each agent's execution complexity is $\mathcal{O}\left(J * E * \left(\sum_{p=2}^{P-1}\left(M_{p-1}^a Z_p^a + Z_p^a Z_{p+1}^a\right)\right)\right)$. The algorithm's convergence analysis considers UAV agents' interaction, influencing subsequent user agent actions while supporting policy changes among agents.

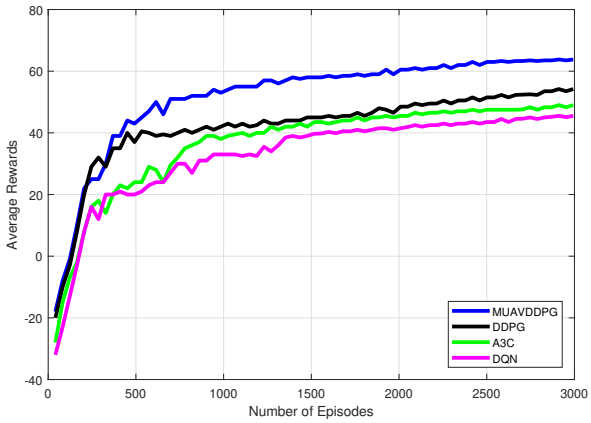


FIGURE 3: Average Rewards with Training Episodes.

V. SIMULATIONS RESULTS AND DISCUSSIONS

In this section, we analyze the outcomes of our algorithm's simulations across various parameter configurations. Our simulations were executed in a Python 3.7 environment utilizing Tensorflow 2.0. The computations were performed on a system powered by a Core i7 CPU operating at 2.4GHz and equipped with 16GB RAM. We consider our analysis to involve a range of 2 to 12 UAVs and 10 to 60 mobile users unless specified otherwise. These UAVs are positioned at an altitude of 100m above ground level. Within our devised multi-agent algorithm, we established 3000 episodes during the training phase, each comprising 200 steps. The neural network employed in this algorithm is a fully connected architecture featuring both a critic network and an actor-network. Specifically, every agent within the system operates with a neural network structure comprising two hidden layers in the actor as well as critic neural networks. These layers consist of 256 neurons in the initial hidden layer and 128 neurons in the subsequent layer. A mini-batch size of 256 was employed to manage the learning process. Other simulation parameters are summarized in Table 2.

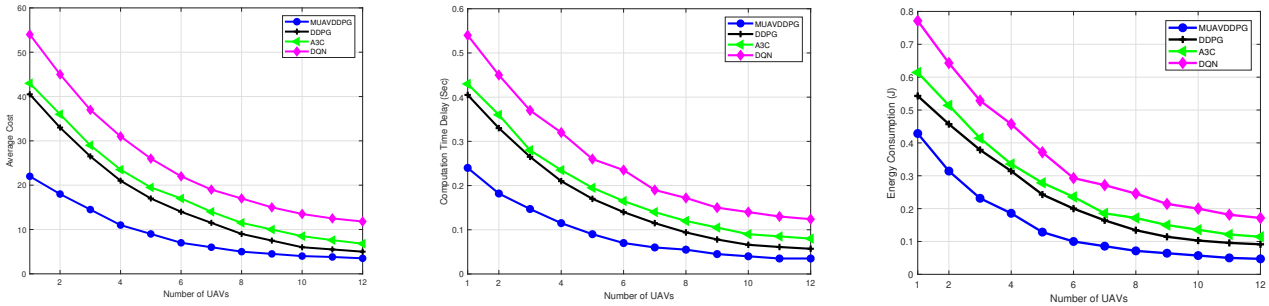
TABLE 2: Experimental Parameters.

Parameters	Value
Task Size	[150-1500] KB
CPU cycles required for task	[150-2000] Mcycles
Maximum processing time delay	[0.1-2] Sec
Mobile Transmission Power of	0.1 W
Task Size	[150-1500] KB
Bandwidth	40 MHz
Learning rate of actor-network	$1e^{-4}$
Learning rate of critic-network	$3e^{-4}$

The proposed MUAVDDPG algorithm is evaluated against the following baseline algorithms.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022



(a) Average Cost (b) Average Time Delay (c) Energy Consumption

FIGURE 4: Effects of UAVs on Cost, Time, and Energy Consumption

1) Deep Deterministic Policy Gradient (DDPG) [28]: A DDPG-based scheme wherein each agent learns from prior offloading experiences and dynamically fellows the nearest computational UAV node.
2) A3C [28]: Both the A3C and DDPG algorithms belong to the paradigm of model-free DRL and follow an actor–critic framework. However, their distinct advantages come to light in various applications, especially in achieving efficient convergence during training.
3) DQN [29]: DQN exhibits strengths in dynamically allocating computational tasks among heterogeneous resources. DQN learns from experience and navigates high-dimensional state spaces contributing to effective decision-making in dynamically changing network conditions.

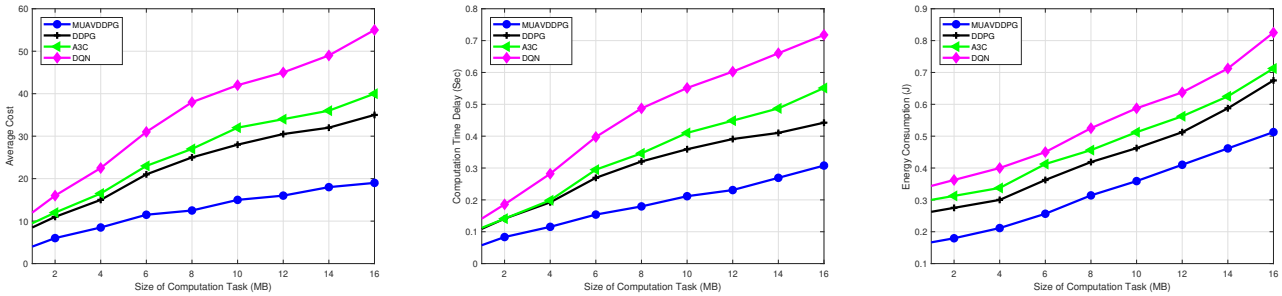
The convergence analysis depicted in Fig. 3 illustrates the performance of various DRL algorithms. Specifically, the proposed MUAVDDPG algorithm demonstrates superior performance compared to the three baseline algorithms. Convergence for the MUAVDDPG algorithm occurs before the 550-episode mark, while the other three algorithms converge around the 700-episode range. This points to the successful alignment of interests between UAVs and mobile users, effectively maximizing rewards and minimizing computation costs through optimal offloading decisions. An observed trend reveals an increase in average rewards proportional to both the number of episodes and the number of agents. Remarkably, the proposed MUAVDDPG algorithm yields an overall reward increase of 16%, 26%, and 33% compared to DDPG, A3C, and DQN, respectively. This notable enhancement in reward values signifies the efficacy of the MUAVDDPG approach in optimizing both UAV and mobile users' objectives within a dynamic network environment.

In Fig. 4, our focus initially centers on analyzing the performance of the proposed algorithm concerning the escalating number of UAVs. As observed in Fig. 4a, a trend emerges, indicating a decrease in the average computation cost across all algorithms with an increase in the number of UAVs. Notably, the MUAVDDPG algorithm showcases a swifter reduction in average cost, surpassing the performance of the three baseline algorithms. This pattern suggests that under the MUAVDDPG framework, agents collaborate more effectively, leading to reduced resource consumption compared to other algorithms. DDPG and A3C exhibit superior performance to DQN due to their enhanced support for distributed learning. Particularly noteworthy is the MUAVDDPG's remarkable reduction in average cost by 30%, 49%, and 70% compared to DDPG, A3C, and DQN, respectively. This demonstrates the MUAVDDPG algorithm's effectiveness within a multi-agent system, surpassing baselines by optimizing agents' decisions. Additionally, Fig. 4b and 4c portray a notable trend: as the number of UAVs increases within the MUAVDDPG algorithm, optimal time delay and energy consumption decrease. This trend is also observed, to a lesser extent, in DDPG, A3C, and DQN. Thus, across various numbers of UAVs, the MUAVDDPG algorithm consistently outperforms the three baseline algorithms by efficiently minimizing time delay and energy consumption.

In Fig. 5, our investigation delves into the average cost, time delay, and energy consumption across various data sizes from 1 to 16 MB. This analysis aims to assess how offloading data sizes impact system performance. Fig. 5a graphs the average cost across diverse data sizes. The observed gradual increase in the average cost across all algorithms is attributed to larger offloading data sizes demanding more time and energy consumption. Tasks with increased data sizes require higher computational time and resources from UAVs. Once the offloading task size reaches 4 MB, there's a rapid spike in the average cost for the three benchmark algorithms. However, under the proposed algorithm, the average cost escalates significantly slower. The algorithm's collaborative nature facilitates information sharing among agents, enabling efficient cost minimization. Comparatively, the proposed algorithm demonstrates a significant reduction in the average cost, showcasing a decrease of 45%, 52%, and 65% when compared to the DDPG, A3C, and DQN algorithms, respectively. Moving to Fig. 5b and 5c, the analysis focuses on time delay and computation energy consumption, encompassing the transmission and execution of mobile users' tasks. Fig. 5b demonstrates a proportional increase in computation time delay with rising offloading data sizes. However, the proposed algorithm consistently maintains lower time delays compared

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022



(a) Average Cost (b) Average Time Delay (c) Energy Consumption

FIGURE 5: Effects of data size on Cost, Time, and Energy Consumption

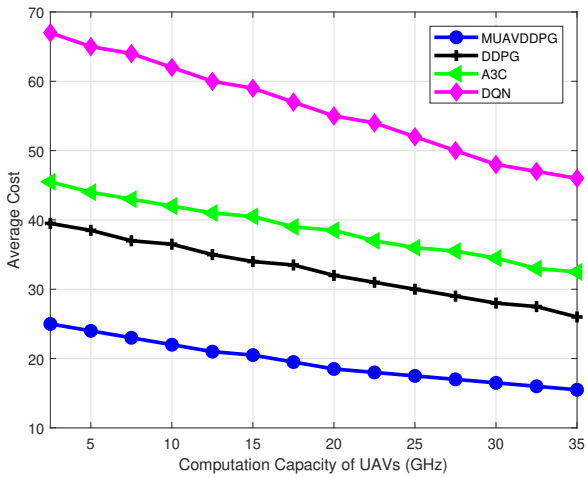


FIGURE 6: UAV Computation Capacity vs Average Cost

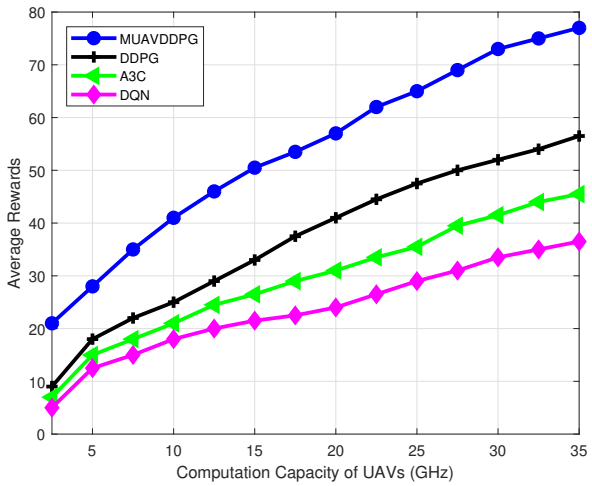


FIGURE 7: UAVs Capacity vs Average Reward

to the three baseline algorithms. Meanwhile, Fig. 5c reveals a general uptrend in total energy consumption across all algorithms as offloading tasks expand. Remarkably, the proposed algorithm exhibits a notably slower increase in energy consumption compared to the three baseline algorithms. This highlights the algorithm's capacity to effectively minimize computation costs concerning time delay and energy consumption across varying offloading data sizes.

The research analyzes how the computational capabilities of UAVs impact the computation cost mobile users incur when offloading tasks. As indicated in Fig. 6, an evident pattern emerges: as the computational capacity of UAVs rises, there is a gradual decline in average costs observed across all algorithms. This decline in average costs demonstrates a consistent, albeit slow, reduction. To elaborate further, it is important to note that the average cost associated with the three fundamental algorithms exceeds that of the proposed MUAVDDPG algorithm. Upon closer examination by evaluating each algorithm's performance concerning the maximum computation capacity, substantial reductions in average costs are evident. Specifically, when comparing the MUAVDDPG algorithm to the DDPG, A3C, and DQN algorithms, there is a notable decrease in average costs by 40%, 52%, and 66%, respectively. This comparison underscores the significant efficacy of the MUAVDDPG algorithm in significantly reducing costs compared to the established algorithms studied.

Fig. 7 and Fig. 8 demonstrate how the computation capacity of UAVs influences the performance metrics of various benchmark algorithms in multi-UAV-assisted MEC environments. Specifically, Fig. 7 illustrates the relationship between UAVs' computation capacity and the average reward achieved by each algorithm, highlighting the increasing trend in average rewards as the computation capacity grows. Notably, the MUAVDDPG algorithm shows the most significant performance enhancement, suggesting that it is particularly effective at utilizing increased computational resources to maximize rewards. This outperformance is consistent across varying capacities, indicating MUAVDDPG's robustness in different computational scenarios. Moreover, Fig. 8 explores the impact of UAVs' computation capacity on the offloading rate, which is a critical measure of how effectively computing tasks are transferred from mobile devices to UAVs. It can

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022

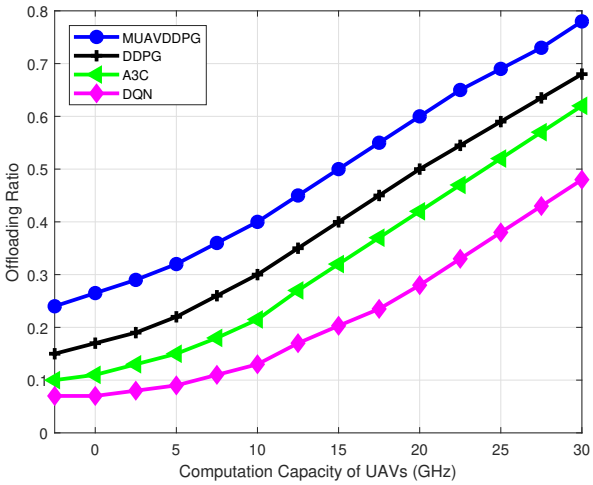


FIGURE 8: UAVs Capacity vs Offloading Rate

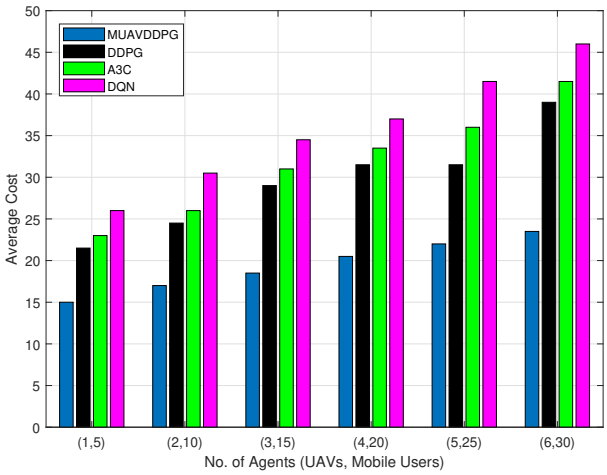


FIGURE 9: Varying No. of UAVs and Mobile Users vs Average Cost

be observed from Fig. 8 that all algorithms benefit from increased computation capacities, achieving higher offloading rates. The superior performance of MUAVDDPG in achieving higher offloading rates further confirms its enhanced ability to efficiently manage and utilize UAV computational resources, which is crucial for task offloading in complex environments. The consistency of MUAVDDPG's superior performance in both average reward and offloading rates under varying computational capacities supports its stability and robustness, making it a promising option for optimizing UAV-assisted operations in diverse MEC settings.

As the demand from mobile users for resources from UAVs rises, so does the average cost. Collaboration intensifies within the ever-changing network setting when there's a growth in the number of agents - including UAVs and mobile users. This collaboration leads to improved rewards for UAVs and cost reductions for mobile users. To confirm the scalability of our proposed framework, we varied the number of agents in our scenario, ranging from 1 UAV and 5 mobile users to 6 UAVs and 30 mobile users across all algorithms, as depicted in Fig. 9. The proposed MUAVDDPG outperforms the three alternative algorithms. Our observations revealed that average costs also escalate as the number of agents rises. However, the cost associated with the MUAVDDPG algorithm remains lower than the others and sees a more gradual increase. Under MUAVDDPG, agents exhibit higher cooperation and a more favorable experience in contrast to DDPG, A3C, and DQN algorithms as the number of participants grows. Hence, in our proposed scenario, the MUAVDDPG algorithm exhibits greater scalability than the other benchmark algorithms and proves its relevance in multi-agent system scenarios.

A. PRACTICAL IMPLEMENTING AND EVALUATING OF THE PROPOSED ALGORITHM

The evaluation of the proposed MUAVDDPG algorithm against various practical scenarios and benchmark approaches reveals its efficacy in optimizing resource allocation and minimizing computation costs in dynamic network environments. Compared to baseline algorithms such as DQN, A3C, and DDPG, the MUAVDDPG algorithm demonstrates superior performance, as illustrated by convergence analysis and average reward increases. Notably, the MUAVDDPG algorithm converges faster, aligning the interests of UAVs and mobile users to maximize rewards and minimize computation costs through optimal offloading decisions.

Furthermore, across different network scales, including varying numbers of UAVs and offloading data sizes, the MUAVDDPG algorithm consistently outperforms baseline algorithms by efficiently minimizing time delay, energy consumption, and average cost. This performance advantage is attributed to the collaborative nature of the MUAVDDPG framework, which facilitates effective information sharing among agents, reduces resource consumption, and ensures optimal resource allocation.

Additionally, the scalability of the MUAVDDPG algorithm is demonstrated by its ability to maintain lower average computation costs even as the number of agents, including UAVs and mobile users, increases. Overall, these findings underscore the practical utility and generalizability of the proposed algorithm in multi-agent system scenarios, highlighting its potential to address complex task offloading and resource allocation challenges in dynamic network environments.

VI. CONCLUSIONS

In this paper, we explored task offloading and resource allocation within a resource-constrained multi-UAV-assisted MEC network. Our focus centered on optimizing compu-

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022

tation costs, encompassing energy consumption and computation delay, by enabling mobile users to access computational resources from UAVs. Leveraging an optimization problem modeled on dynamic preferences and employing the MDP framework, our goal was to minimize computational demands for mobile users.The MUAVDRL algorithm effectively addresses this challenge by efficiently navigating its dynamic and high-dimensional nature. Rigorous simulation results validate our proposed framework's superior performance and effectiveness when compared to existing state-of-the-art methods, highlighting its substantial potential in practical scenarios. Notably, the simulation outcomes underscore the efficacy of the MUAVDRL algorithm, showcasing its ability to significantly enhance system utility while meeting stringent latency and energy requirements, surpassing the capabilities of other baseline algorithms.

REFERENCES

[1] S. Raza, S. Wang, M. Ahmed, M. R. Anwar et al., "A survey on vehicular edge computing: architecture, applications, technical issues, and future directions," Wireless Communications and Mobile Computing, vol. 2019, 2019.

[2] S. Raza, W. Liu, M. Ahmed, M. R. Anwar, M. A. Mirza, Q. Sun, and S. Wang, "An efficient task offloading scheme in vehicular edge computing," Journal of Cloud Computing, vol. 9, pp. 1–14, 2020.

[3] Z. Yu, Y. Gong, S. Gong, and Y. Guo, "Joint task offloading and resource allocation in uav-enabled mobile edge computing," IEEE Internet of Things Journal, vol. 7, no. 4, pp. 3147–3159, 2020.

[4] M. A. Mirza, Y. Junsheng, S. Raza, M. Ahmed, M. Asif, A. Irshad, and N. Kumar, "Mcla task offloading framework for 5g-nr-v2x-based heterogeneous vecns," IEEE Transactions on Intelligent Transportation Systems, 2023.

[5] M. A. Mirza, Y. Junsheng, M. Ahmad, S. Raza, W. U. Khan, F. Xu, and A. Nauman, "Drl-driven zero-ris assisted energy-efficient task offloading in vehicular edge computing networks," Journal of King Saud University-Computer and Information Sciences, p. 101837, 2023.

[6] M. Ahmed, S. Raza, M. A. Mirza, A. Aziz, M. A. Khan, W. U. Khan, J. Li, and Z. Han, "A survey on vehicular task offloading: Classification, issues, and challenges," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 7, pp. 4135–4162, 2022.

[7] X. Huang, X. Yang, Q. Chen, and J. Zhang, "Task offloading optimization for uav-assisted fog-enabled internet of things networks," IEEE Internet of Things Journal, vol. 9, no. 2, pp. 1082–1094, 2021.

[8] Z. Chen, Z. Chen, Y. Jia, and L.-C. Wang, "Residual energy-aware caching in energy harvesting-based mobile d2d network," IEEE Transactions on Green Communications and Networking, vol. 4, no. 3, pp. 845–857, 2020.

[9] J. Lu, S. Wan, X. Chen, Z. Chen, P. Fan, and K. B. Letaief, "Beyond empirical models: Pattern formation driven placement of uav base stations," IEEE Transactions on Wireless Communications, vol. 17, no. 6, pp. 3641–3655, 2018.

[10] H. Cao, W. Zhu, Z. Chen, Z. Sun, and D. O. Wu, "Energy-delay tradeoff for dynamic trajectory planning in priority-oriented uav-aided iot networks," IEEE Transactions on Green Communications and Networking, vol. 7, no. 1, pp. 158–170, 2022.

[11] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot et al., "Mastering the game of go with deep neural networks and tree search," nature, vol. 529, no. 7587, pp. 484–489, 2016.

[12] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in International conference on machine learning. PMLR, 2016, pp. 1928–1937.

[13] M. A. Mirza, J. Yu, S. Raza, M. Krichen, M. Ahmed, W. U. Khan, K. Rabie, and T. Shongwe, "Drl-assisted delay optimized task offloading in automotive-industry 5.0 based vecns," Journal of King Saud University-Computer and Information Sciences, p. 101512, 2023.

[14] X. Chen, H. Zhang, C. Wu, S. Mao, Y. Ji, and M. Bennis, "Performance optimization in mobile-edge computing via deep reinforcement learning," in 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall). IEEE, 2018, pp. 1–6.

[15] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for v2v communications," IEEE Transactions on Vehicular Technology, vol. 68, no. 4, pp. 3163–3173, 2019.

[16] F. Sun, N. Cheng, S. Zhang, H. Zhou, L. Gui, and X. Shen, "Reinforcement learning based computation migration for vehicular cloud computing," in 2018 IEEE Global Communications Conference (GLOBECOM). IEEE, 2018, pp. 1–6.

[17] S. Zhu, L. Gui, N. Cheng, Q. Zhang, F. Sun, and X. Lang, "Uav-enabled computation migration for complex missions: A reinforcement learning approach," IET Communications, vol. 14, no. 15, pp. 2472–2480, 2020.

[18] M. Min, L. Xiao, Y. Chen, P. Cheng, D. Wu, and W. Zhuang, "Learning-based computation offloading for iot devices with energy harvesting," IEEE Transactions on Vehicular Technology, vol. 68, no. 2, pp. 1930–1941, 2019.

[19] N. H. Motlagh, M. Bagaa, and T. Taleb, "Energy and delay aware task assignment mechanism for uav-based iot platform," IEEE internet of things journal, vol. 6, no. 4, pp. 6523–6536, 2019.

[20] S. Jeong, O. Simeone, and J. Kang, "Mobile edge computing via a uav-mounted cloudlet: Optimization of bit allocation and path planning," IEEE Transactions on Vehicular Technology, vol. 67, no. 3, pp. 2049–2063, 2017.

[21] T. Zhang, Y. Xu, J. Loo, D. Yang, and L. Xiao, "Joint computation and communication design for uav-assisted mobile edge computing in iot," IEEE Transactions on Industrial Informatics, vol. 16, no. 8, pp. 5505–5516, 2019.

[22] X. Huang, C. Peng, Y. Wu, J. Kang, W. Zhong, D. I. Kim, and L. Qi, "Joint interdependent task scheduling and energy balancing for multi-uav enabled aerial edge computing: A multi-objective optimization approach," IEEE Internet of Things Journal, 2023.

[23] U. Saleem, Y. Liu, S. Jangsher, X. Tao, and Y. Li, "Latency minimization for d2d-enabled partial computation offloading in mobile edge computing," IEEE Transactions on Vehicular Technology, vol. 69, no. 4, pp. 4472–4486, 2020.

[24] F. Zhou, Y. Wu, R. Q. Hu, and Y. Qian, "Computation rate maximization in uav-enabled wireless-powered mobile-edge computing systems," IEEE Journal on Selected Areas in Communications, vol. 36, no. 9, pp. 1927–1941, 2018.

[25] J. Chen, Q. Wu, Y. Xu, N. Qi, T. Fang, L. Jia, and C. Dong, "A multi-leader multi-follower stackelberg game for coalition-based uav mec networks," IEEE Wireless Communications Letters, vol. 10, no. 11, pp. 2350–2354, 2021.

[26] S. Raza, M. Ahmed, H. Ahmad, M. A. Mirza, M. A. Habib, and S. Wang, "Task offloading in mmwave based 5g vehicular cloud computing," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 9, pp. 12 595–12 607, 2023.

[27] M. A. Ebrahim, G. A. Ebrahim, H. K. Mohamed, and S. O. Abdellatif, "A deep learning approach for task offloading in multi-uav aided mobile edge computing," IEEE Access, vol. 10, pp. 101 716–101 731, 2022.

[28] A. M. Seid, G. O. Boateng, S. Anokye, T. Kwantwi, G. Sun, and G. Liu, "Collaborative computation offloading and resource allocation in multi-uav-assisted iot networks: A deep reinforcement learning approach," IEEE Internet of Things Journal, vol. 8, no. 15, pp. 12 203–12 218, 2021.

[29] I. Khan, S. Raza, R. Khan, K. Nahida, X. Tao et al., "A deep learning-based algorithm for energy and performance optimization of computational offloading in mobile edge computing," Wireless Communications and Mobile Computing, vol. 2023, 2023.

MUHAMMAD NAQQASH TARIQ obtained his Bachelor's degree (BEE) from The University of Faisalabad (TUF), Faisalabad in 2012, and his Master's degree in Electronics & Communication Engineering from the Beijing University of Posts and Telecommunications, Beijing, China in 2017. Currently, he is a Ph.D. candidate at the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include Unmanned Ariel vehicular communication networks, Software Defined Networking, fog and edge computing, internet of things, machine learning and deep reinforcement learning, and big data analytics.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2024.3411022

JINGYU WANG (Senior Member, IEEE) received the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2008. He is currently a Professor with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications. He has published more than 100 articles in international journals, including IEEE Communications Magazine, IEEE TRANSACTIONS ON CLOUD COMPUTING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. His research interests include the IoV and AIoT, SDN, overlay networks, and traffic engineering.

SALMAN RAZA received the MCS degree from Bahauddin Zakariya University, Multan, Pakistan, with a distinction (Gold Medal) in Computer Science in 2009, the M.S degree from the G.C. University, Faisalabad, Pakistan, in Computer Science in 2014. He has completed his Ph.D. degree from the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China. He is currently working as an Assistant Professor at National Textile University Faisalabad, Pakistan. His current research interests include mobile edge computing, vehicular networks, task offloading, and deep reinforcement learning.

MOHAMMAD SIRAJ (Senior Member IEEE) received his Ph.D. from Universiti Technologi Malaysia. He holds Master of Engineering degree in Computer Technology and Applications from Delhi College of Engineering, Delhi, India and Bachelor of Engineering degree in Electronics and Communication Engineering from Jamia Millia Islamia, New Delhi. He has worked as a Scientist in Defense Research and Development Organization, India. Currently he is working as an Assistant Professor in Electrical Engineering, King Saud University. His research interests are Cognitive Wireless Networks, Wireless Mesh Networks, Sensor Networks, Internet of Things, Cloud Computing and Telecom Optical Networks. He has numerous peered publications in well-known international journals and Conferences. He is a Senior Member IEEE and reviewer of many well-known International Journals and Conferences.

MAJID AL TAMIMI (Member, IEEE) received the B.Sc.Eng. degree (Hons.) in electrical engineering from King Saud University, Saudi Arabia, in 2004, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Waterloo, Canada, in 2010 and 2014, respectively. In 2015, he joined the Department of Electrical Engineering, King Saud University, as an Assistant Professor. He served as a TA with the Department of Electrical and Computer Engineering, University of Waterloo, in 2013, and the Department of Electrical Engineering, King Saud University, from 2004 to 2006. His current research interests include analyzing the energy cost for wireless handheld devices and cloud computing architecture, integrating mobile computing with cloud computing, and studying and designing green ICT solutions.

DR. SAIFULLAH MEMON received a master's degree from Quaid-e-awam University of Engineering, Science and Technology, (QUEST) in 2015. Onward he acquired his Ph.D. degree from Beijing University of Posts and Telecommunications (BUPT), China in 2023 in the field of communication engineering. He is the author of more than 30 research articles and is currently working as an Assistant Professor at QUEST. His research interests include wireless communication, sensors, wireless body area networks (WBAN), QoS issues in sensor and ad-hoc networks, routing algorithms, and network security.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4

International Journal of Intelligent Transportation Systems Research
https://doi.org/10.1007/s13177-024-00403-6

Train Length and Service Frequency Optimization: Passenger Health Expense Perspective

Manish Kumar Sharma[1,2,3] · **Manvendra Singh**[4,5,6,7]

Received: 1 March 2024 / Revised: 9 May 2024 / Accepted: 10 May 2024
© The Author(s), under exclusive licence to Intelligent Transportation Systems Japan 2024

Abstract
The urban transport rail network is represented by operational lines and associated variables such as the origin/destination matrix, ridership, and speed. In railway planning, operators always desire to maximize profit by evaluating the effects of different train lengths and service frequencies, whereas passengers aspire to minimize travel expenses. This study analyses the effects of train length and service frequency on economic expenses related to passengers' travel expenses and the critical impact on the health costs incurred due to acute stress. An optimization model for maximizing the operator's profit and minimizing passenger travel expenses, along with a health costs model related to travel expenses, is presented. A simplified algorithm is developed to analyse the impact of train length variation on the infrastructure cost, which is tested and compared over a set of real implemented projects. The scenarios of short trains, high frequency trains over long trains, and low service frequency are analysed. It is observed that longer trains with lower service frequencies optimize the operator's profit and passenger travel expense for intercity travel. The findings indicate that selecting a balance factor of approximately 0.6 is crucial for maintaining optimal passenger expenses. Moreover, within the range of 0.7 to 0.9, passenger benefits reach a peak while slightly compromising the operator's profit. It is evident that the 9-car train length with a service frequency of 16 and an optimum headway of 15 min, over a 4-hours observation duration, yields superior benefits for both passengers and operators. The effectiveness of the optimization model is further analysed with respect to the operator's performance, which is adversely affected by keeping the frequency and train size constant during the peak and off-peak periods.

Keywords Train length · Service frequency · Optimum profit · Passenger travel expense · Travel stress · Nonlinear optimum model

1 Introduction

With the speed and scale of urbanization, accelerated population and economic activities have generated a potential rise in the demand for rapid, affordable, and efficient metro transportation systems. Leveraging benefits such as rapid, clean, and efficient ways to transport passengers through urban metro transportation systems has relieved the traffic pressure on many clusters of metropolitan cities. The policy dimensions of the urban rail transportation network include the length of lines, the distance between stopping stations, service pricing, the operating frequency of trains, and train length. Passenger arrival flow is the key factor in making decisions about these policy dimensions for a public transport operator. Specifically, during peak hours of any working day, the passenger arrival flow is enthusiastically large. During such a large flow improvement, the main objective of planning for the operator is to improve the total travel time, maintain headway regularity and minimize passengers' generalized travel expenses [1].

With large policy dimensions, a vast research area is available about public transportation planning. In the areas of network design, crew management, trip demand evaluation, etc., issues have been addressed through many approaches. In metro cities with painful traffic congestion, parking may lead car drivers to opt for public transport if

Highlights
- *Impact analysis of train length and service frequency on the rail-based infrastructure.*
- *Impact analysis of travel expenses mapped to acute stress affecting passenger health.*
- *Model for optimizing train length and service frequency to minimize travel expenses and maximize operator profits.*
- *A long train length with a low frequency is needed for optimal intercity travel.*

Extended author information available on the last page of the article

Published online: 05 June 2024

an affordable and service-quality system is offered. In [2], the author performed a cluster analysis through a customer satisfaction survey on mass transit rail in Tehran in which satisfaction with public transit was influenced by trip characteristics and sociodemographic groups. *Structural equation modelling* (SEM) was utilized to identify the most effective attributes in each group. The results showed that both clustering and segmentation based on individuals' characteristics were efficient. Rail transport operators are responsible for planning and providing transport systems that maximize the quality of service to passengers and ensuring minimum *operation and maintenance* (O&M) costs by adopting novel initiatives in an urban rail transit system. In one of the rail-based urban transit systems of China, for service quality improvements, the author analysed operational strategies such as optimizing coach composition, energy consumption, and relative traction capacity [3]. Similarly, in [4], the author tried to ameliorate the train schedule and timetable rationalization by exploring human intervention and adopting innovative ways of matching supply with demand.

The user desires rapid rail transport services to be cost friendly, comfort providers for train and transfer terminals, and punctual; to have broader service coverage; and to have a higher frequency level. The operator, on the other hand, thrives on maximizing profit and plans to employ larger trains [5]. However, planning with only this policy dimension will be unfavorable to passengers, leading to low frequencies, high schedule delay costs, more waiting times, and additional time losses for boarding and alighting. The planning objective for rail transport is to obtain an equilibrium between the objectives of these two conflicting stakeholders [6, 7].

Equilibrium or optimization of these two objectives can be achieved by a two-step process in which the first step is line planning, which includes selecting and constructing terminal stations, tracks, and trains [8–10]. The service frequencies of the trains are planned in the second step. Mostly, the actions are planned jointly, but because the passenger flow is time dependent and elastic, service frequency planning needs to be performed for different periods or demand changes. The demand variations trigger the need for modification in the frequency and length or composition of trains.

Traditionally, longer trains with reasonable service frequencies have been introduced to meet the objective of meeting the elastic passenger demand of rail operators, including high-speed railway systems worldwide. Globally, networks of high-speed rail and intercity rapid rail transit systems have manifolded, so the service frequency and train length choices are planned, and test cases are available. The choice of train length and service frequency has a considerable impact on the operator's infrastructure and operating costs; hence, there is considerable research interest in this optimization challenge.

Researchers have presented different optimization methodologies [11] and maximizing the number of direct trips has been found to be the most common objective [12]. Headway flexibility and the augmentation of line capacity to increase the attractiveness of rail networks were also considered [8, 13]. Primarily, the solutions are focused on minimizing the passenger travel expense mapped to the waiting and other comfort-related factors.

The rail transportation sector and passenger behavior are directly related to our fast-paced world engrossed by microwave speed culture in daily life. Our complete life has become deadline-oriented, and all tiny activities are sequential and linked to another activity in such a manner that if the deadline of one activity is missed, then the deadline for completing the next sequence of activities collapses.

Apart from generating high acute stress due to putting off a task until its deadline, researchers have also found dead brain cells, depleted creativity, and debilitating effects on one's health due to deadlines or timeline-based living [14]. Urban or metropolitan transportation is primarily linked with passengers travelling for work and normally treats travelling as a task to be completed during a scheduled duration or at a scheduled time. After the completion of this task, the main important task related to work is taken up, which sets up the context for the developed acute stress of honoring phased deadlines. Consequently, the task of urban travel can be termed a rush-recurring deadline, which is critically linked with the completion of all the timelines of subsequent activities. Hence, for urban rail transport passengers, travel expenses include direct fare or ticket costs and detrimental health costs due to acute stress because of waiting time, schedule delay costs and congestion. Therefore, in addition to mapping economic aspects, passenger travel expenses must essentially be associated with consequential health cost expenditures [15, 16].

1.1 Related Work: Choice of Service Frequency and Train Length

The rapid metro rail system is a vital transit mode in modern metropolises and has attracted significant attention from researchers in recent decades. The literature on urban rail transport includes two main categories: rail line planning, e.g., line length, number of stations, train length, etc., and the service frequency of trains. These studies have focused primarily on the train timetabling problem, aiming to determine a preoperation schedule or train rescheduling model that needs to adjust the preoperative timetable effectively to counter the effects of any disturbances [17].

The urban rail passenger desires to minimize the total travel expense [18], and at the same time, the operator wants to maximize the profit, which leads to an optimization challenge between the choice of train length and the service

frequency. In [19], the author addressed the effect of train length and service frequency on the determination of train timetables, where the service strategies of different operators, such as short trains with high frequency and long trains with low frequency, were analysed, in which a maximum profit model with elastic demand was solved using an analytic method. Early work [20] proposed a nonlinear dynamic programming model with headway and passenger costs as two dynamic equations. The model presented an iterative hierarchical multilevel decomposition method. This study focused primarily on headway rescheduling; i.e., service frequency referenced to fluctuations in passenger demand and passenger travel expense was not considered. In [21], the authors presented an optimization model using dynamic and stochastic transit assignment that accounted for variations in headways and crowding and their consequences for passenger flow distribution. The supply-setting problem is formulated as a system cost minimization problem consisting of service users' (generalized travel) costs and service providers' (fixed and variable) costs. In [22], the authors optimized the quality of service with the mean time spent by the passengers and safety constraints on the headway of trains as variables. The case of nonsmooth arrival rates of passengers was focused on headway optimization. The authors in [23] proposed a linear programming-based model predictive control formulation methodology for computing the optimal train schedules. The methodology could generate optimal schedules for the whole day and between two separate periods.

Real-time scheduling and routing of trains were implemented through an algorithm, and several theoretical and practical results were discussed [24]. In [25], the authors discussed train schedules that minimize the total passenger waiting time by utilizing a nonlinear integer programming model with linear constraints. A mixed-integer linear programming approach was used to design dynamic passenger demand-oriented metro train scheduling to minimize energy consumption and waiting time [26]. A mixed-integer nonlinear programming model using a novel linearization approach was also used to explain the train service configuration and passenger assignment problems to minimize operators' costs, passengers' waiting time costs and transfer costs [27]. The same optimization model was also used for Beijing URT line capacity utilization and trip time reduction without additional operating costs and investments. The results revealed that the number of transfers can be reduced by reducing travel time and operating costs when the crossline express delivery operating mode is properly adjusted [28].

Different methods for setting service frequency and train length were presented to address the trade-off between operational and passenger travel costs [29–31]. Most analyses were performed for a line with two stations and were generalized to the network. The different lengths of stations and passenger flow variabilities at different stations were not considered. In [32], the author explained the relationship between high-speed train service level and passenger flow variation with the help of the logit and gravity model on the Beijing-Shanghai high-speed rail line by classifying the stations in the large, medium and small station categories based on ridership; the results revealed that large station origin–destination (OD) should reduce train stops and increase travel speed overall to attract more transferring passengers; and medium and small station OD should increase train stops and increase service frequency to attract more potential passengers.

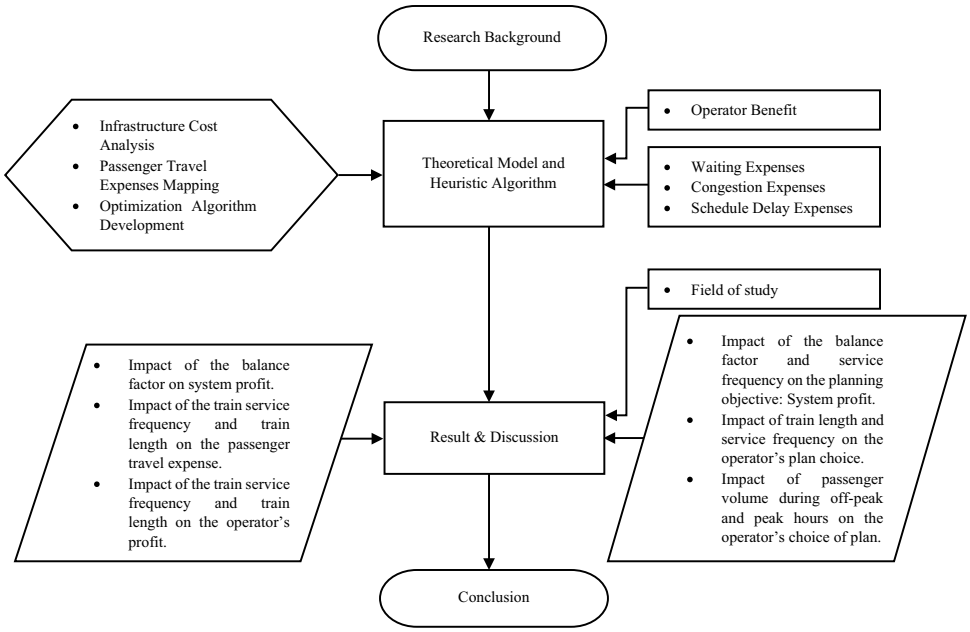
A similar optimization issue was addressed in [33], where a model was proposed for solving the trade-off between the service frequency and the number of coaches per train. In [34], a model was formulated in which elastic passenger demand and social benefits were maximized by using train length, service frequency and cost as control variables. The empirical data showed that length and frequency change approximately with the square root of demand. In [35], the authors considered two forms of schedule delay: one with very high frequencies and passengers generally ignoring the timetables and another where timetables are consulted. It was observed that two types of schedule delay have different effects on choice service frequency. The authors proposed an optimization model based on the behavior of passengers when choosing which train to board [36]. The relationship between the train operation frequency and train formation length was shown to be an s-continuous p-median problem, and the train scheduling optimization problem was simplified [37].

Although the traditional models discussed above simplify the optimization of trade-offs faced in capacity management in rail operations, the effect on the operator's benefit through variations in the infrastructure cost of rail networks, referenced to the variation in train length and frequency, and mapped to passenger travel expenses, e.g., waiting time, schedule delay, and congestion, on the health costs incurred by the passenger and a simplified optimization algorithm to analyse the impact of train length and service frequency on the overall system profit are open issues.

To the best of our knowledge, a novel approach comprising train length and frequency impact on an operator's infrastructure cost, health cost mapping of passenger travel expenses and simplified optimization analysis of complete lines with different passenger arrival flows at different stations elevated or *underground* (UG) is illustrated in this paper.

The objectives of this research primarily involve steps carried out in the study as demonstrated in Fig. 1. The remainder of this paper is structured as follows: Section 1 presents the research background and related works on the service frequency choice and train length. Section 2

Fig. 1 Research workflow



formulates the proposed theoretical model and heuristic algorithm to optimize the operator's profit and passenger generalized travel expense for variations in the size and frequency of trains. In Section 3, analyses, results and discussion are presented based on empirical data from the *Regional Rapid Transit System* (RRTS) project of the *National Capital Region Transport Corporation* (NCRTC). The conclusions are presented in Section 4.

2 Theory and Method

This research paper focuses on optimizing train length and service frequency to increase overall system profit. This involves minimizing passenger travel expenses while maximizing the operator's profit. The train length, which represents the number of coaches, and service frequency play crucial roles in optimizing railway schedule planning at the tactical level, which is one of the three hierarchical levels of time interval-based schedule planning, alongside strategic and operational procedures. The primary objectives and novel contributions of this paper are outlined as follows:

- **Infrastructure Cost Analysis:** The variation in train length and frequency affects infrastructure costs, including factors such as viaducts, tunnels, and platform lengths. To analyse these effects, this study introduces a novel method that assesses how train length and frequency impact infrastructure costs, which are pivotal in evaluating operator profits.
- **Passenger Travel Expenses Mapping:** Passenger travel expenses encompass waiting times, schedule delays, and congestion-related costs. This paper delves into a comprehensive analysis of these expenses and proposes models to connect them with the health costs borne by passengers.
- **Optimization Algorithm Development:** To comprehensively examine the influence of train length and service frequency on the overall system profit, this research presents an innovative optimization algorithm. This algorithm enables a thorough evaluation of how changes in these parameters impact the entire system's financial performance.

Through these contributions, this research enhances our understanding of optimizing train operations by considering both infrastructure costs and passenger experiences, ultimately leading to improved profitability for operators.

2.1 Modelling the Objective

A novel optimization algorithm for the RRTS network plan is proposed with two objectives: minimizing the passengers' total travel expense and maximizing the operator's profit. The effects of train length and service frequency on the overall system profit are analysed. The "overall system profit" pertains to the combined financial outcome of the Rapid Rail Transit System (RRTS) network plan. This outcome is influenced by two primary objectives: reducing the total travel expenses incurred by passengers and increasing the profit earned by the operator. The effects of train length and service frequency on this overall system profit are then examined. To enable the demonstration of the fundamental

concepts without loss of generality, the following primary considerations are made in this paper.

- One of the objectives of this paper is to observe the relationship between train length and service frequency; hence, the ticket price is assumed to be fixed, although passenger flow is partially related to the fare component of overall travel expenses.
- To reveal the relationship between train length and service frequency in terms of system profit, the network is assumed to be linear with multiple stations in between, and the distance between two adjacent stations is treated as a line for modelling, as shown in Fig. 2.
- For a complete network, the profit is maximized using all analyses for all lines cumulatively. As the RRTS network progresses through different stations between cities, the passenger flow for each station varies. The passenger flow for each station is assumed to be the same as that projected through the survey conducted during the detailed project planning phase.
- The passenger's comprehensive travel expenses encompass various factors, including the fare cost and the health-related costs associated with acute stress or tight deadlines that arise from the physical and psychological toll of prolonged waiting times, the anxiety and pressure resulting from schedule disruptions, and the physical discomfort and mental strain endured due to congested travel conditions. In essence, the passenger's total travel expense calculation considers not only the direct monetary costs but also the less tangible and equally significant health-related expenses incurred during the journey.
- The observation period (t_{obp}) is assumed to be a chosen period within which the intensity of passenger demand is unified at a specific station. This is taken as the peak period where the passenger density is the maximum during running hours per day.
- The train service frequency and length are chosen in such a manner that the capacity generated during the observation duration must be sufficient to allow all passengers to arrive during that period.

Under these assumptions, the effects of the ticket cost, planned stopping schedule and variations in demand on the train service schedule are ignored. Hence, the relationship between train length and service frequency on system profit can be effectively revealed. The urban transport system desires planning objectives to minimize passengers' total travel expenses ($p_{totaltravexp}$) and maximize total operator profits (op_{prft}). The two conflicting requirements are transformed into a single objective of maximizing system profit (sys_{prft}) by introducing the balance factor (bal_{fact}), which confirms a balance between the operator's profit and the passengers' total travel expenses. The objective is simplified to maximize (sys_{prft}) with a determination of the number of trains (trn_n) and the length of the train (trn_l) for an observation duration (t_{obp}). Thus, the optimization model of system profit for the RRTS network is formed as in Eq. (1):

$$sys_{prft} = bal_{fact} \text{x} p_{totaltravexp} + \left(1 - bal_{fact}\right) \text{x } op_{prft} \tag{1}$$

2.2 Operator Benefit

The time fluctuation characteristic of the passenger volume of the railway network is echoed in the planning horizon by determining the train schedules for each specified observation duration (t_{obp}). From the perspective of urban rail transport, an operator's benefit is primarily derived from ticket income. When the price of a ticket is (p_{tkt}) or the passenger flow is (p_{vol}) during the observation duration (t_{obp}), the operator's benefit (op_{bnft}) is derived as per Eq. (2):

$$op_{bnft} = p_{tkt} \text{ x } p_{vol} \tag{2}$$

By providing train services to passengers, the cost incurred by the operator can be divided into two types: one infrastructure and another operation cost. Hence, the operator's cost is the sum of the infrastructure cost and O&M cost.

The infrastructure requirement needs modification with the variation in train length. The fixed infrastructure required to operate the RRTS rail network comprises different subsystems, namely, Civil Infrastructure, Track, General Electrical, *Overhead Equipment* (OHE) in Electric Traction, *Signalling*

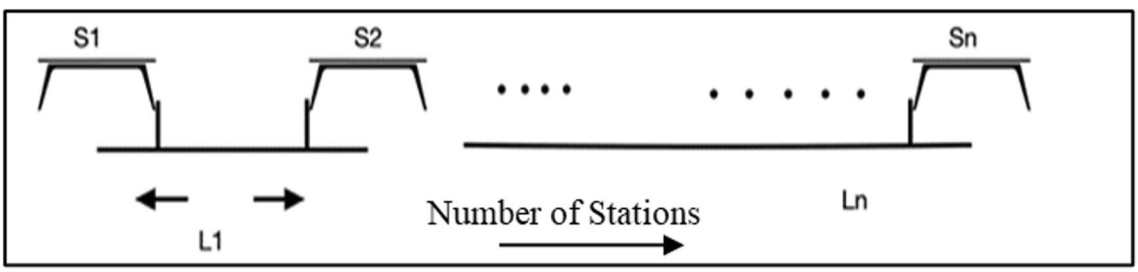


Fig. 2 RRTS station and line network

and Telecommunication (S&T), and *Platform Screen Door* (PSD). The civil infrastructure may further be dependent upon the variation in platform length, viaduct and tunnel volume for the station area referenced to the variation in required train length. Similarly, the straight track length, length of sidings and terminals, *Inspection Bay Line* (IBL), and *Stabling Bay Line* (SBL) required for maintenance and depot length also vary. This impact occurs because, for shorter trains, such as those with the same length as mentioned above for sidings, the IBL is also shortened, which significantly affects the cost component. The cost impact of such variations in the civil and track subsystems echoes that of the general electrical, S&T cable and traction subsystems. The effect on the PSD system is consequently apparent owing to the lower requirement of doors for shorter trains and consequential provision of infrastructure, such as cabling.

The RRTS rail network is taken as a linear network with several stations from start to end. The starting end of the network assumes station S_1 to be the first station, followed by S_2 and the end station S_n. The distance between the stations is also not uniform; hence, the length of the rail network between two stations is taken as lines L_1, $L_2 \ldots L_n$ for pairs of stations S_1—S_2, $S_2 - S_3$, … $S_{n-1} - S_n$, as shown in Fig. 2.

Apropos of the above considerations, an *infrastructure cost determination algorithm* (ICDA) with pseudocode is designed as explained in Table 1 below to determine the infrastructure cost required for running a specific train-length service.

As the rail network is continuously put into passenger service for approximately 80% of the day, the system needs dedicated maintenance to re-energize for the next day. Hence, the operator must incur maintenance expenditures apart from operation expenditures, e.g., crew, station staff, and power logistics. Concerning variations in the length of trains and service frequency operation costs related to crew management and energy requirements for running trains are significant variables. As the infrastructure cost varies for trains of different lengths, the maintenance efforts required also vary; however, coach maintenance is considered a nonvariable assuming that the total number of coaches remains the same only when the length of trains and service frequency are optimized within this constraint. From the above deliberations, the *O&M cost determination algorithm* (OCDA)

Table 1 ICDA

```
Input:
{
   location: [S1, S2, S3, ..., Sn],
   unit-cost: Infra subsystems,
   unit coach length: u_cl,
   no_ug_stn: no_ug_stn
}

//Define platform lengths for three different trains
platform_lengths: [pl_3, pl_6, pl_9]

//Calculate total platform length for each iteration
total_platform_length: []
for each pl_iteration in platform_lengths:
   total_platform_length.append(pl_iteration x 1.5 x no_ug_stn)

//Calculate total track length for each iteration
total_track_length: []
for each pl_iteration in platform_lengths:
   total_track_length.append(sum(N_L, depot_length x fct_iteration, terminal_length x fct_iteration))

//Map all the above lengths to total units corresponding to all subsystems
total_units_iteration: []
for i in range(len(platform_lengths)):
   tunl_iteration = total_platform_length[i]
   vdctl_iteration = total_platform_length[i] - total_platform_length[(i + 1) % len(platform_lengths)]
   trkl_iteration = total_track_length[i]

   total_units_iteration.append(map{tunl_iteration, vdctl_iteration, trkl_iteration, N_L})

//Calculate Infrastructure cost for each iteration
Infra_cst_iteration: []
for i in range(len(platform_lengths)):
   Infra_cst_iteration.append(sum(mul{unit-cost: Infra subsystems}, total_units_iteration[i]))

End

Output: { Infra_cst_iteration }
```

with pseudocode is formulated as explained in Table 2 below to obtain the O&M cost required for running a specific train length and with a particular service frequency.

Considering trains of length (trn_l) with a service frequency of (trn_n) should suffice to carry a load of passengers in the (t_{obp}) duration for the lines so that no passengers are left out, the infrastructure cost determined by ICDA, *Infra_cst_iteration* needs to be mapped to the duration of the study (t_{obp}). Assuming that the operator creates the infrastructure for running trains to the minimum possible headway for a fixed number of years before heavy replacement is needed, the mapping of the infrastructure cost is performed in reference to the number of trains running during the observation duration. Similarly, the O&M cost, *O&M_cost_iteration,* is mapped to the duration of the study (t_{obp}). Hence, the operator's profit is determined from Eq. (3):

$$op_{prft} = op_{bnft} - trn_n \text{x} \left| infra_{cost_{iteration}} \right|_{t_{obp}} + trn_n \text{x}\ trn_l \text{ x}\ unit_{coach_{cost}} + O\&M_cost_iteration \quad (3)$$

2.3 Passenger's Travel Expense

The traditional approach of considering the time spent by passengers in transportation may involve allocation to other activities and has pecuniary value. As urban transport systems have higher speeds and service frequencies are also high, the financial mapping of the waiting time, travel time, etc., is small in comparison to the actual fare expenses.

However, urban rail travel can be categorized as a rush-recurring deadline task that directly impacts the completion of all the timelines of subsequent activities planned by the passenger. Hence, for urban rail transport passengers, this rush-recurring-deadline task can cause acute stress due to waiting time, schedule delay costs and congestion. Therefore, in addition to economic valuation, passenger travel expenses must essentially be considered in reference to the substantial health cost expenditures incurred due to developed acute stress. The passenger's travel expenses due to travel time, congestion expense and schedule delay expense are modelled to map this cost in terms of the health expenditure incurred by the passenger.

2.3.1 Waiting Expenses

The monetary mapping of travel expenses such as waiting time, scheduled departure time or congestion expenses may not be directly perceived or appreciated by the passenger due to the high service frequency, i.e., less headway of RRTS. The health costs referenced in the above travel expenses are real perceptible expenses incurred by the passengers whom they observe during the entire engagement period with RRTS. During the waiting time, acute stress develops during the state of health, as passengers may need to reach a place at a certain time, and then waiting for travel will lead to an escalated level of anxiety transformable into acute stress. Subsequently, the passenger's body reacts by releasing hormones, which makes the passenger's brain more alert, causing the muscle to tense and increasing the heart rate and blood pressure, which vary depending upon how much waiting time the passenger is observing. If the waiting time in terms of headway considered (15 min in the present case) increases from 0 to 80%, the acute

Table 2 OCDA

```
Input:
{
   {trn_n, trn_l, t_obp},
   {unit-cost: coach_maintenance per coach per hour, uc_cm},
   {all subsystems maintenance cost per unit network length per hour, alsbsys_mntce_iteration},
   {energy cost: per unit network length per train, enrg_cst_iteration},
   {crew cost: per hour per train, crew_cst_iteration}
}

For iteration: 3 to 9 in steps of 3:
do
  //O&M cost for specific train length
   O&M_cost_iteration: sum{
      uc_cm x trn_l x trn_n x t_obp,
      alsbsys_mntce_iteration x l1,
      alsbsys_mntce_iteration x l2,
      alsbsys_mntce_iteration x ln,
      enrg_cst_iteration x trn_n x l1,
      enrg_cst_iteration x trn_n x l2,
      enrg_cst_iteration x trn_n x ln,
      crew_cst_iteration x trn_n
   }
End

Output: {O&M_cost_iteration}
```

stress shown in Fig. 3 does not affect passengers much; however, during the remaining 20% of headway, a sharp increase in acute stress puts the passenger at risk for health problems, in turn increasing the travel expense, i.e., the health cost.

The phenomenon effect of deadline rush is described by a hyperbolic function [38], as shown in Eq. (4):

$$with_{wt} = \frac{without_{wt}}{1 + k\text{x } wt} \tag{4}$$

where $with_{wt}$ is the cost after waiting time, wt is exhaust, $without_{wt}$ is the base expenditure without any waiting time, wt is the waiting time, and k is the passenger-dependent parameter. The higher k is, the more impulsive the passenger is, and the steeper the curve is. For research analysis, the mean k value is taken as 0.1 [39].

For a mass transit system, as the headway is lower than that for long-distance mainline trains, the passenger arrival may be assumed to be uniformly distributed during the waiting period. For the RRTS system, trains run on a fixed headway; thus, the waiting time (wt) is related to the service frequency and cost after waiting, as given in Eq. (5):

$$with_{wt} = \frac{without_{wt}}{1 + \frac{k\text{x } p_{distribution}\text{x } t_{obp}}{trn_n}} \tag{5}$$

where $p_{distribution}$ is the passenger distribution parameter and, for a uniform distribution, is taken as 0.5. Considering the base expenditure as unity, Eq. (5) is simplified, and the total cost referenced to the waiting time is determined as in Eq. (6):

$$cst_{wt} = \int_0^{wt} \frac{1}{1 + k\text{x } t} dt \tag{6}$$

$$cst_{wt} = \frac{1}{k} \ln\left(k\text{x } \frac{p_{distribution}\text{x } t_{obp}}{trn_n}\right) \tag{7}$$

Fig. 3 Acute stress and waiting time

2.3.2 Congestion Expenses

During the observation duration (t_{obp}), a fixed number of identical passengers opts for the line of the RRTS network to work. When a train approaches the station, all passengers on the platform thrive to board it so that their waiting time expense is minimized. The passengers of a train encounter the expected crowding discomfort ($ds_{cmfrt(np)}$), where np is the number of users taking the same train. Discomfort is caused by additional passengers encroaching on the minimum area (approx. 1 m2) occupied by passengers for comfortable travel [40]. The function $ds_{cmfrt(np)}$ is strictly increasing for approximately two-thirds of the total number of stations for intercity travel because at stations close to the end of the network, passengers start deboarding the train. The discomfort function is modelled as in Eq. (8):

$$ds_{cmfrt(np)} = \frac{p_{vol}\text{x } t_{obp} - trn_n\text{x } trn_l\text{x } cc_c}{trn_n\text{x } trn_l\text{x } c_l\text{x } c_w} \tag{8}$$

The coach carrying capacity ($\boldsymbol{cc_c}$) is taken as 180 for comfortable travel, and the available area per coach is taken as a product of coach length ($\boldsymbol{c_l}$) 23.54 m long) and coach width ($\boldsymbol{c_w}$) 3.24 m wide coach, ignoring the area occupied by fittings, etc., of the coach. The congestion cost function is minimal and approximately zero at the origin station when the train is empty. Travelling in crowded trains is associated with increased mental acute stress since the passenger perceives less control over the environment. The impact on mental health outweighs that of other urban acute stressors, such as pollution and noise [41].

The acute stress developed due to congestion increases slowly at the initial stations and then increases sharply immediately after a couple of stations, as shown in Fig. 4.

The acute stress can be modelled as in Eq. (9):

$$stress_{congestion} = (ds_{cmfrt(np)})^2 \tag{9}$$

Therefore, the total congestion expense for the complete network during the observation duration t_obp is given by Eq. (10):

$$congestion_{cost} = \sum_1^n (ds_{cmfrt(np)})^2 \tag{10}$$

2.3.3 Schedule Delay Expenses

In an ideal scenario, the departure times of trains must coincide with the passengers' expected boarding times to ensure that the rail network is suitable for passengers. However, trains run with a fixed headway in urban rail transport, and there is a difference between the passenger's desired boarding time and the operator's provided boarding options [42]. If the passenger arrives early to

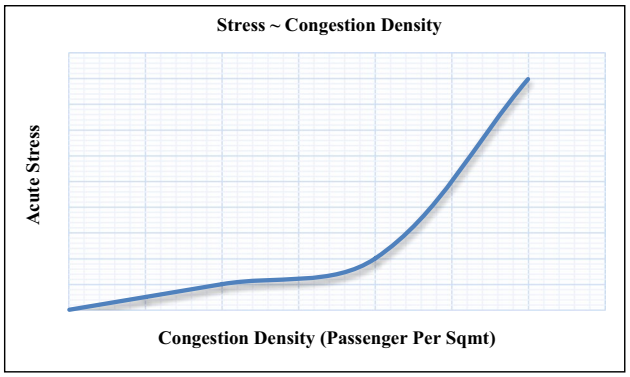


Fig. 4 Acute stress and congestion density

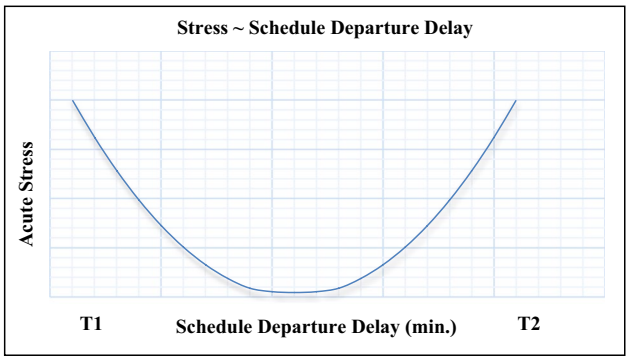


Fig. 5 Acute stress and schedule departure delay

the arriving train (T2), there is an anxiety that he should have arrived a little later and may have completed other tasks apart from exactly matching the train time. This spike in anxiety is reflected in the development of acute stress, which affects the health costs incurred. Similarly, if the passenger arrives a little late to the train that has just departed (T1), then the acute stress is increased to the level that passengers analyse all previous events and tasks in a short time, exposing themselves to a sudden spike of higher anxiety, which is shown in Fig. 5.

This anxiety stress is an expense to the passenger and follows Eq. (11):

$$sde_{stress} = \begin{cases} \frac{1}{e^{\frac{t_{obp}}{trn_n}}}, 0 < pass_{arrival_{time}} \leq \frac{t_{obp}}{2*trn_n} \\ e^{\frac{t_{obp}}{trn_n}}, \frac{t_{obp}}{2*trn_n} < pass_{arrival_{time}} \leq \frac{t_{obp}}{trn_n} \end{cases} \tag{11}$$

Assuming that the passenger flow between headways is uniformly distributed, the total schedule delay cost can be determined as in Eq. (12):

$$sde_{cost} = p_{distribution} \text{ x} \left\{ \int_0^{\frac{t_{obp}}{2\text{x } trn_n}} \frac{1}{e^{\frac{t_{obp}}{trn_n}}} d(trn_n) + \int_{\frac{t_{obp}}{2*trn_n}}^{\frac{t_{obp}}{trn_n}} e^{\frac{t_{obp}}{trn_n}} d(trn_n) \right\} \tag{12}$$

Using Taylor series expansion and ignoring the minimal effect of higher-order terms on sde_{cost}, the schedule delay cost may be approximated as in Eq. (13):

$$sde_{cost} = p_{distribution} \text{x } t_{obp} \text{x } trn_n^{-2} \tag{13}$$

Considering the above discussion, a passenger's total travel expense, which consists of the fare, waiting time, congestion expense and schedule delay expense, can be determined by the *passenger expense determination algorithm* (PEDA) with pseudocode, as explained in Table 3.

The passenger's total travel expense is determined by combining the ticket cost, waiting expense, schedule delay expense and congestion expense. Costs other than the ticket cost are mapped in financial terms with transformation factors *q1*, *q2*, and *q3* for the waiting expense,

Table 3 PEDA

```
Input:
{
   {cst_wt, 〖congestion〗 _cost, sde_cost},
   {p_distribution, t_obp},
   {mapping factor: q1, q2, q3}
}

For iteration: trn_n and trn_l in different steps
//Iteration steps should ensure the capacity generated, trn_n x trn_l x capacity_coach must be sufficient to
carry all passengers in the observation duration t_obp
do
   p_tr_expns_ite: p_tkt + q1 x cst_wt +  q2 x congestion_cost + q3 x sde_cost)
End

Output: {p_totaltravexp: p_tr_expns_ite}
```

schedule delay expense and congestion expense, respectively, and expressed as in Eq. (14):

$$p_{totaltravexp} = p_{tkt} + q1 \times cst_{wt} + q2 \times congestion_{cost} + q3 \times sde_{cost} \quad (14)$$

The planning objectives of the operator can be achieved by minimizing the passenger's total travel expense ($p_{totaltravexp}$) and maximizing the total operator's profit (op_{prft}) determined through the deliberations mentioned above.

3 Results & Discussion

The Delhi-Ghaziabad-Meerut (NCRTC) RRTS corridor is approximately 82 km in length and has 23 scheduled stoppages between two Intercity trips; this corridor is used for the analysis of optimization issues, as shown in Fig. 6 [43]. The observation duration *t_obp* is taken as the window of peak or rush hours of 4 h. length from the daily running hours. During these rush hours, the passenger flow is at its maximum, and all the passenger expenses are at their peak, which leads to the fair assumption that at other times, the optimization solution will reduce the passenger travel expense compared to that of the peak hour analysis.

The length of the viaduct, platform, depot, tunnel, sidings and other relevant infrastructure parameters affected by the train length of the 9 coaches are explained in Table 4. Similarly, parameters for other train length parameters are also determined and put in as input to the ICDA algorithm for the calculation of infrastructure cost.

The per unit cost of the above variables for the 9-car train length is also determined from the real implemented project data in the metro rail sector in India and referenced from the Report on Operations and Maintenance Systems for Metro Railways, Ministry of Urban Development, and formulated in Table 5. The capacity of the coach is taken as 180 (no. of passengers) for a standard gauge network, and the coach length is 23.54 mt with a width of 3.24 mt. The unit of expense and system profit is taken in the INR, and the transport demand is taken in the number of passengers [44].

The balance factor (*bal_fact*) is varied from 0 to 1 in steps of 0.1 for comprehensive analysis. The effects of train length and service frequency on system profit, operator profit and passenger travel expense with respect to variations in the balance factor are analysed and discussed.

3.1 Impact of the Balance Factor on System Profit

The system profit (in INR) decreases sharply as the balance factor increases because for higher values of the balance factor, the objective is aligned toward maximizing the operator's profit, and passengers' travel expenses are not given due weightage. The effect of the system profit on the balance factor is shown in Fig. 7.

A set of ten service frequencies based on the minimum and maximum headways planned are analysed, and it is observed that when the service frequency is higher, the rate of decrease in system profit is lower, providing minor compensation for passenger expenses. Thus, a very high value of the balance factor should not be chosen to maintain the passenger expenses at an optimum level.

3.2 Impact of the Train Service Frequency and Train Length on the Passenger Travel Expense

Figure 8 shows the effect of train service frequency on passenger travel expenses relative to intercity travel. As the train service frequency (number of trains running in an hour) increases, the passenger travel expense decreases. Additionally, longer trains with high frequency are inferred to offer maximum benefits to passengers as the congestion cost is minimized along with other travel expenses. However, a short train with a high service frequency provides similar benefits to passengers as does a long train with such a high service frequency. The mode of a short train with a higher frequency is suitable for the case of abundant transport capacity and high service frequency demand, whereas the operation cost is minimized for long trains with less frequency.

3.3 Impact of the Train Service Frequency and Train Length on the Operator's Profit

As shown in Fig. 9, with an increase in service frequency, a decrease in the operator's profit is observed for a given train length. This observation is in line with the fact that as the service frequency increases, the operator must deploy additional system resources to maintain and run the trains at such a frequency. The reduction in the operator's profit is greater for the 9-car train than for the 3-car train, which indicates that a large train formation with less service frequency can improve the operator's profit.

3.4 Impact of the Balance Factor and Service Frequency on the Planning Objective: System Profit

The balance factor ensures a balance between the operator's train running cost and the passenger travel expense. Figure 9 illustrates the effect of the balance factor on the system benefit concerning a particular service frequency. The passenger benefit is more significant for a higher balance factor, and the operator's profit is reduced. In

Fig. 6 NCRTC Delhi-Ghaziabad-Meerut RRTS Corridor

general, the region of interest shown in Fig. 10 describes the rational range of the balance factor observed between 0.4 and 0.6, where the decrease in system profit concerning the service frequency maintains the maximum system profit. If the balance factor is set higher in the range from 0.7 to 0.9, the passenger benefits are maximized, whereas the operator incurs a heavy reduction in profit. Although this approach ensures more benefits to passengers, this network operation is sustainable only if the operator is supported through enough financial subsidies from the government.

3.5 Impact of Train Length and Service Frequency on the Operator's Plan Choice

The operator's decision about train length and frequency is generally related to the plan's profitability for running trains. The influence of train length and frequency are embodied in

Table 4 Infrastructure details

Station Name	Station Code	Viaduct Length (m)	No of Platform at station (No.)	Platform length (m)	Tunnel length (m)	Track/sidings length (m)	Depot length (m)
Sarai Kale Khan	S0	352.08	6	215	-	1994.16	-
New Ashok Nagar	S1	5187.352	2	215	-	10804.704	-
Anand Vihar	S2	6610.523	2	290.09	-	13801.226	-
Sahibabad	S3	6026.091	2	215.6	6608.397	25700.176	-
Ghaziabad	S4	4943.88	3	215.6	-	10534.56	-
Guldhar	S5	4233.089	2	215.6	-	8897.378	-
Duhai	S6	3861.89	2	215.6	-	8154.98	-
Duhai Depot	S7	2020.5	2	215	-	4471	-
Murad Nagar	S8	6048.33	2	215	-	12526.66	11618.88
Modi Nagar South	S9	6710.712	2	215	-	13851.424	-
Modi Nagar North	S10	3890.399	2	215	-	8210.798	-
Meerut South	S11	8205.385	2	215	-	16840.77	-
Partapur	S12	2674.614	2	-	-	5349.228	-
Rithani	S13	1516.289	2	-	-	3032.578	-
Shatabdi Nagar	S14	2018.058	2	215	-	4466.116	-
Brahmpuri	S15	1892.281	2	-	-	3784.562	-
Meerut Central	S16	1360.123	1	-	-	2720.246	-
Bhaisali	S17	2128.382	1	-	-	4256.764	-
Begumpul	S18	1129.877	1	215	-	2474.754	-
MES Colony	S19	2108.081	2	-	-	4216.162	-
Daurli	S20	2709.5	2	-	-	5419	-
Meerut North	S21	1601.18	2	-	-	3202.36	-
Modipuram	S22	1962.32	2	216	-	4356.64	-
Modipuram Depot	S23	3348.024	2	-	-	6696.048	11618.88

Table 5 Unit cost

Item	Unit	Cost/unit (INR in Cr)
Viaduct	meter	0.060
Platform	meter	0.375
Tunnel	meter	0.196
Track	meter	0.011
Depot Yards/IBL/SBL	meter	0.028
OHE Cost UG	meter	0.008
OHE Length of wire	meter	0.005
S&T Equipments & Cables	meter	0.008
PSD Equipments & Cables	per door	1.690
Rolling Stock Cost	per coach	7.850
Annual Maintenance Cost (Rolling Stock including Manpower)	per coach per hour	253.290
Annual Maintenance Cost (Other Systems including Manpower)	per meter per hour	0.054
Energy Consumption Cost	per train per meter	0.199
Operations Manpower Cost	per train per hour	570

the objective function, which is described as system profit. The effects of train length and service frequency are illustrated in Fig. 11.

When the train service frequency is increased, the system profit increases and approaches nearly the same values at very high service frequencies. When the train length and service frequency are very small, the operator's cost for running trains is low, the service capability is also low, and the service quality is poor. With increased service frequency and train length, passengers' travel expenses are reduced, so operators' profit also decreases. However, for the operator's selection, the region of interest shown in Fig. 11 is impacted by the frequency in the range of 10 to 20, which provides the maximum system profit. Within those frequencies of interest, a 9-car train length with a service frequency of 16 optimizes the benefits for both the operator and passengers. When the observation duration is considered to be 4 h, the optimum headway observed is 15 min with a 9-car train, for which the operator's profits are maximized and the passenger travel expense is minimized.

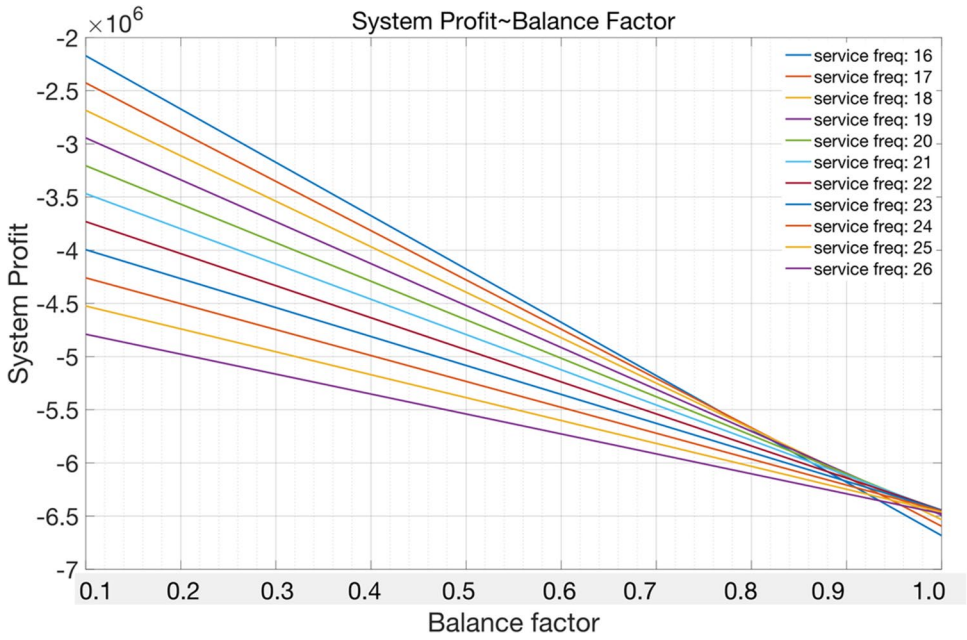


Fig. 7 System profit and balance factors

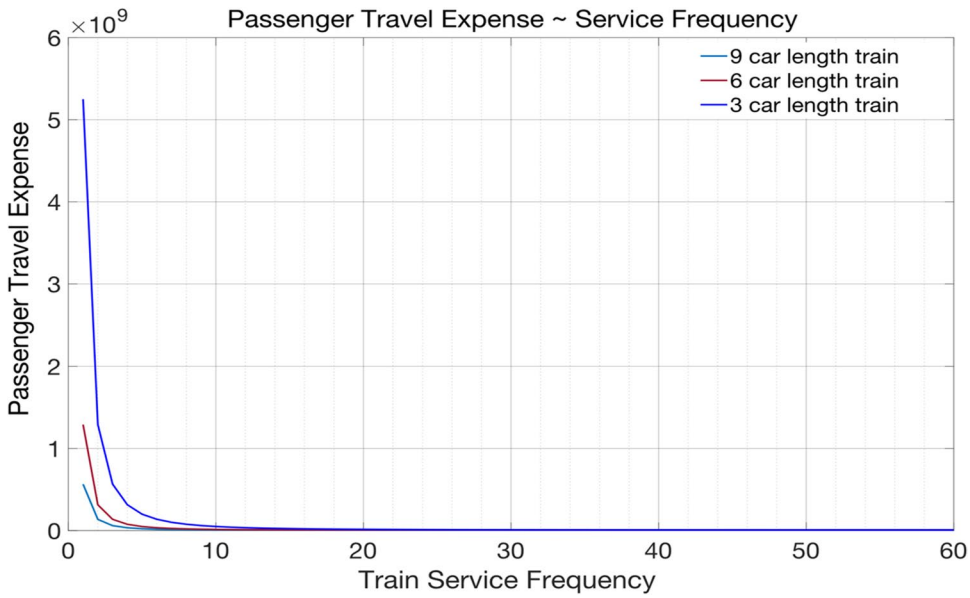


Fig. 8 Passenger travel expense and service frequency (per hour) and train length

3.6 Impact of Passenger Volume During Off-Peak and Peak Hours on the Operator's Choice of Plan

The passenger volume reaches a maximum during the rush hours when the day starts, and most passengers desire to reach the workplace; similarly, during the evening, they desire to reach home from the workplace. However, during the middle hours of the day or early start, the passenger flow is comparatively low compared to that during rush or peak hours. These durations are termed off-peak hours. In general, the operator's decision about train length and frequency should cater to the available passengers with acceptable system profits and hence need to be dynamic in accordance with the passenger flow. The effects of off-peak and peak-hour passenger volumes are illustrated in Fig. 12.

The off-peak hour system profit approaches the peak hour system profit for a specific train service frequency. Notably, in comparison to the peak-hour system profit, the off-peak-hour system profit approaches the same values for lower service frequencies. Hence, if the operator continues to maintain the same service frequency and train size during off-peak and peak hours, the operator's performance will be adversely affected. For the operator's selection, the region of interest, as shown in Fig. 12, observes that a frequency in

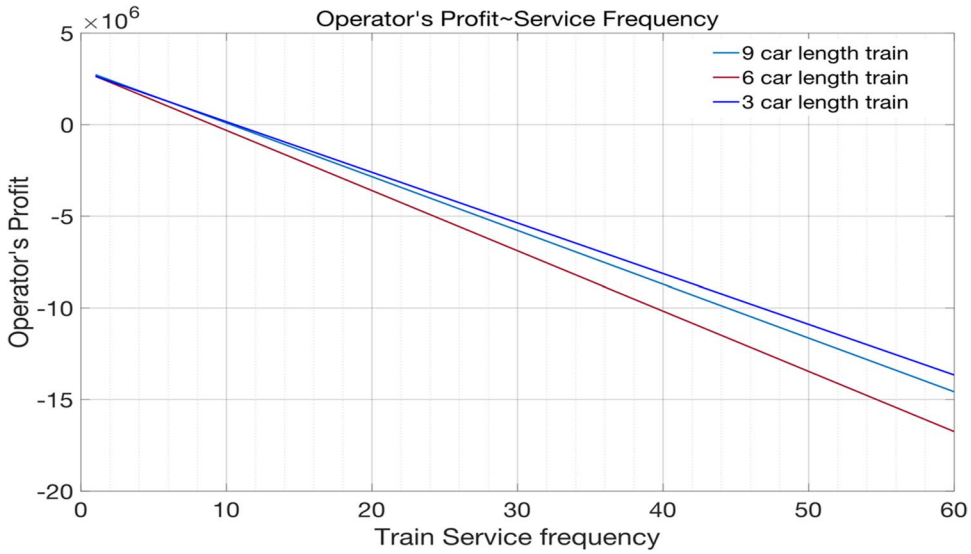


Fig. 9 Profit and service frequency (per hour) and train length of the operator

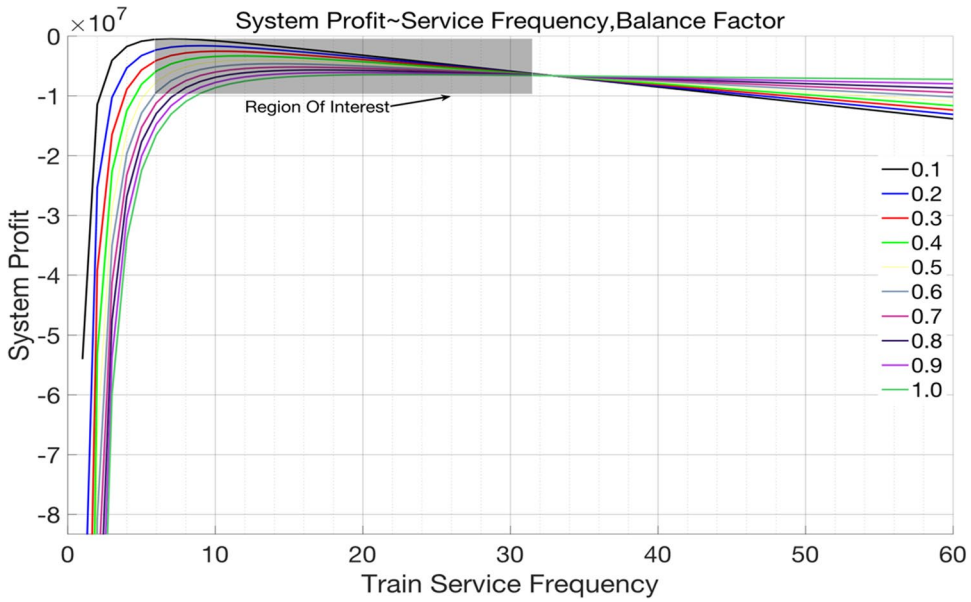


Fig. 10 System profit and service frequency (per hour) and balance factor

the range of 5 to 10 provides maximum system profit during off-peak hours compared to a service frequency range of 10 to 20 in peak hours.

4 Conclusion

The objective of the study is to choose the train length and service frequency so that the operator's profit and passenger expense are optimized. The variation in the infrastructure cost of rail networks due to variations in the viaduct, tunnel, and platform length owing to the train length and service frequency is embodied in the analysis. The acute stress that passengers experience due to travel expenses such as waiting time, schedule delays and congestion expenses is mapped. The study analysed and presented acute stress models to map these expenses to the health costs incurred by passengers.

The results of the optimization model reveal that a balance factor of approximately 0.6 should be chosen to maintain the optimal passenger expenses. If the balance factor is considered to be between 0.7 and 0.9, the passenger benefits are maximized on the verge of the operator's reduced profit. The region of interest observes that a service frequency in the range of 5 to 10 provides maximum system profit in off-peak hours compared to a range of 10 to 20 in peak hours. Thus, it is evident that a 9-car length train has a service frequency of 16 and an optimum headway of 15 min for an observation duration of 4 h. optimize the benefits for both the operator and the passengers.

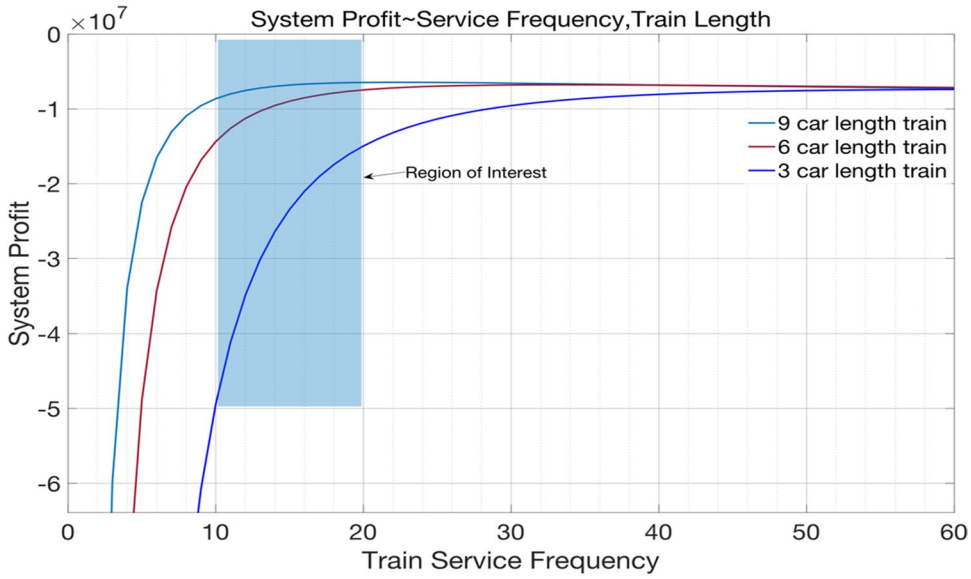


Fig. 11 System profit and service frequency (per hour) and train length

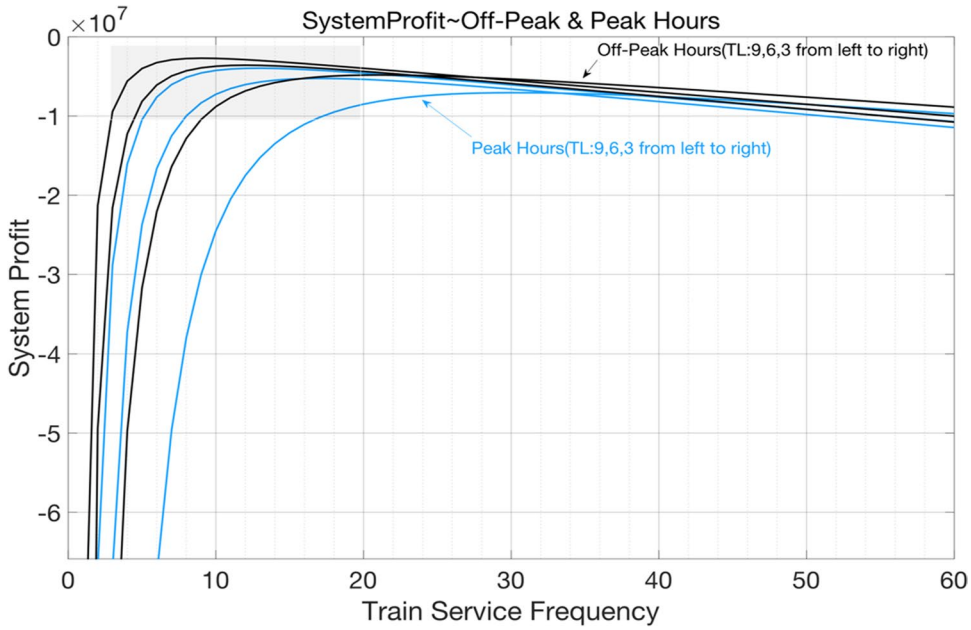


Fig. 12 System profit and service frequency (per hour): Off-peak and peak-hour passenger volumes

The study demonstrated that the model of a short train with a higher frequency is suitable only when transport capacity and high service demand are available and when operators' decision to opt for a long train length with a low frequency is the optimal operation decision for intercity travel. Furthermore, the operator should not opt for a constant train length or service frequency plan during peak and off-peak periods.

Despite the abundant research literature on rail-based transportation, there is still a gap in the study of how the choice of train length and service frequency are correlated with psychological stress. During the research study, both the constraints, ticket price and passenger flow, are considered constant and uniformly distributed, but over time, both of these constraints are variable in nature and require further deliberation. Thus, future research work should include the analysis of train formation and scheduling with the variability factor of these constraints.

Acknowledgements The authors hereby declare that manuscript has not been published before or submitted to any other journal for the publication and acknowledge to abide by all rules and regulations of the Journal.

Authors' Contributions Manish Kumar Sharma (Corresponding author): Visualization, Methodology, Investigation, Writing—Original Draft, Resources, Data Curation, Proofreading.

Manvendra Singh (Coauthor): Conceptualization, Validation, Formal analysis, Writing—Review & Editing.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Yes.

Consent for Publication Yes.

Competing Interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Fernandez, A.: Predictive traffic regulation for metro loop lines based on quadratic programming. Proc. Inst. Mech. Eng., Part F: J. Rail Rapid Transit **220**(2), 79–89 (2006). https://doi.org/10.1243/09544097F00505
2. Mesbah, M., Sahraei, M., Soltanpour, A., and Habibian, M.: Perceived service quality based on passenger and trip characteristics: A structural equation modelling approach. J. Rail Transp. Plan. Manag. **23** (2022). https://doi.org/10.1016/j.jrtpm.2022.100340.
3. Feng, X., Liu, H., Zhang, H., Wang, B., and Sun, Q.: Rational formations of a metro train improve its efficiencies of both traction energy utilization and passenger transport. Math. Probl. Eng. Article ID-643274, 1–7 (2013) https://doi.org/10.1155/2013/643274
4. Sharma, M.K., Chauhan, B. K.: Timetable rationalization & operational improvements by human intervention in an urban rail transit system: An exploratory study. Transp. Res. Interdiscip. Perspect (ISSN: 2590–1982) **13** (2022) https://doi.org/10.1016/j.trip.2021.100526
5. Vickrey, W.: A review of: "Transportation Economics" by Herbert Mohring, Balllnger Publishing Company, Cambridge. Eng. Econ. **23**(3), 205–207 (1977). https://doi.org/10.1080/00137917708913066
6. Guihaire, V., Hao, J.K.: Transit network design and scheduling: A global review. Transp. Res. A: Policy Pract. **42**(10), 1251–1273 (2008). https://doi.org/10.1016/j.tra.2008.03.011
7. Michaelis, M., Schöbel, A.: Integrating line planning, timetabling, and vehicle scheduling: a customer-oriented heuristic. Public Transport. **1**(3), 211–232 (2009). https://doi.org/10.1007/s12469-009-0014-9
8. Schmidt, M.: Integrating routing decisions in public transportation problems. Vol. 89, Schmidt, M. E. Springer New York (2014) https://doi.org/10.1007/978-1-4614-9566-6
9. Nemhauser, P.: Handbooks in operations research and management science. Elsevier (2007)
10. Borndörfer, R., Grötschel, M., Pfetsch, M.E.: A column-generation approach to line planning in public transport. Transp. Sci. **41**(1), 123–132 (2007). https://doi.org/10.1287/trsc.1060.0161
11. Schöbel, A.: Line planning in public transportation: models and methods. OR Spectrum **34**(3), 491–510 (2012). https://doi.org/10.1007/s00291-011-0251-6
12. Bussieck, M. R.: Optimal lines in public rail transport. https://nbnresolving.org/urn:nbn:de:gbv:084-114020 (1998) https://doi.org/10.24355/dbbs.084-200511080100-231. Accessed 21 Dec 1998
13. Albrecht, T.: Automated timetable design for demand-oriented service on suburban railways. Public Transport. **1**, 5–20 (2008). https://doi.org/10.1007/s12469-008-0003-4
14. McEwen, B.S.: Neurobiological and Systemic Effects of Chronic Stress. Chronic Stress (Thousand Oaks) (2017). https://doi.org/10.1177/2470547017692328
15. Key, K.: Stress: the dark side of deadlines. Psychology Today. (2015) https://www.psychologytoday.com/us/blog/counseling-keys/201506/the-dark-side-deadlines. Accessed 15 June 2015
16. Bot, I., Kuiper, J.: Stressed brain, stressed heart. Lancet **389**(10071), 770–771 (2017). https://doi.org/10.1016/S0140-6736(17)30044-2
17. Parbo, J., Nielsen, O., and Prato, C.: Passenger perspectives in railway timetabling: a literature review. Transp. Rev. 1–27 (2016) https://doi.org/10.1080/01441647.2015.1113574
18 Zhu, Z.: Analysing influencing factors of transfer passenger flow of urban rail transit: a new approach based on nested logit model considering transfer choices. Int. J. Environ. Res. Public Health. **18**(16), 8462 (2021). https://doi.org/10.3390/ijerph18168462
19. Deng, L., Zeng, Q., Zhou, W., Shi, F.: The effect of train formation length and service frequency on the determination of train schedules. Proc. Inst. Mech. Eng., Part F: J. Rail Rapid Transit **228**, 378–388 (2013). https://doi.org/10.1177/0954409713478589
20. Cury, J., Gomide, F., Mendes, M.: A methodology for generation of optimal schedules for an underground railway system. IEEE Trans. Autom. Control **25**(2), 217–222 (1980). https://doi.org/10.1109/TAC.1980.1102282
21. Cats, O., Gluck, S.: Frequency and vehicle capacity determination using a dynamic transit assignment model. Transp. Res. Rec. **2673**(3), 574–585 (2019). https://doi.org/10.1177/0361198118822292
22. Minciardi, R., Paolucci, M., and Presenti, R.: 1995. Generating optimal schedules for an underground railway line. In Proceedings of 34th IEEE Conference on Decision and Control. https://doi.org/10.1109/CDC.1995.479246
23. Assis, W.O., Milani, B.E.A.: Generation of optimal schedules for metro lines using model predictive control. IFAC Proceedings Volumes. **35**(1), 475–480 (2002). https://doi.org/10.3182/20020721-6-ES-1901.01547
24. Mannino, C., Mascis, A.: Optimal Real-Time Traffic Control in Metro Stations. Oper. Res. **57**(4), 1026–1039 (2009). (http://www.jstor.org/stable/25614813)
25. Niu, H., Zhou, X., Gao, R.: Train scheduling for minimizing passenger waiting time with time-dependent demand and skip-stop patterns: Nonlinear integer programming models with linear constraints. Transp. Res. B: Methodol. **76**, 117–135 (2015). https://doi.org/10.1016/j.trb.2015.03.004
26. Yin, J., et al.: Dynamic passenger demand-oriented metro train scheduling with energy-efficiency and waiting time minimization: Mixed-integer linear programming approaches. Transp. Res. B: Methodol. **97**, 182–213 (2017). https://doi.org/10.1016/j.trb.2017.01.001
27. Li, Z., Zhao, J., and Peng, Q.: Train service design in an urban rail transit line incorporating multiple service routes and multiple train compositions. Transp. Res. C: Emerging Technol. **123** (2021) https://doi.org/10.1016/j.trc.2020.102959

28. Yang, A., Wang, B., Huang, J., and Li, C.: Service replanning in urban rail transit networks: Cross-line express trains for reducing the number of passenger transfers and travel time. Transp. Res. C: Emerging Technol. **115** (2020). https://doi.org/10.1016/j.trc.2020.102629
29. Ibarra-Rojas, O.J.: Planning, operation, and control of bus transport systems: A literature review. Transp. Res. B: Methodol. **77**, 38–75 (2015). https://doi.org/10.1016/j.trb.2015.03.002
30. Gkiotsalitis, K., Cats, O.: Reliable frequency determination: Incorporating information on service uncertainty when setting dispatching headways. Transp. Res. C: Emerging Technols. **88**, 187–207 (2018). https://doi.org/10.1016/j.trc.2018.01.026
31. Gkiotsalitis, K.: A model for the periodic optimization of bus dispatching times. Appl. Math. Model. **82**, 785–801 (2020). https://doi.org/10.1016/j.apm.2020.02.003
32. Yan, W., Lei, N., and Yu, K.: Evaluation of the impact of high-speed train service level on passenger flow. IOP Conf. Ser.: Earth and Environ. Sci. 820, International Conference on Mechanics and Civil, Hydraulic Engineering, June 2021, Kunming, China. (2021). energy consumption https://doi.org/10.1088/1755-1315/820/1/012028
33. Goossens, J.W., Hoesel, S., Kroon, L.: On solving multitype railway line planning problems. Eur. J. Oper. Res. **168**(2), 403–424 (2006). https://doi.org/10.1016/j.ejor.2004.04.036
34. Oldfield, R.H., Bly, P.H.: An analytic investigation of optimal bus size. Transp. Res. B: Methodol. **22**(5), 319–337 (1988). https://doi.org/10.1016/0191-2615(88)90038-0
35. Jansson, K.: Optimal Public Transport Price and Service Frequency. JTEP **27**(1), 33–50 (1993)
36. Shi, J.: Cooperative passenger flow control in an oversaturated metro network with operational risk thresholds. Transp. Res. C: Emerging Technols. **107**, 301–336 (2019). https://doi.org/10.1016/j.trc.2019.08.008
37. Kondo, R., Shiomi, Y., Uno, N.: Network Evaluation Based on Connectivity Reliability and Accessibility. Transp. Res. Econ. Policy Netw. Reliab. Pract., Springer. **1**, 131–149 (2012). https://doi.org/10.1007/978-1-4614-0947-2_9
38. Konig, C.J., Kleinmann, M.: Deadline rush: a time management phenomenon and its mathematical description relationships between critical thinking and attitudes toward women's roles in society. J. Psychol. **139**(1), 33–45 (2005). https://doi.org/10.3200/JRLP.139.1.33-45
39. Petry, N.M., Kirby, K.N., Kranzler, H.R.: Effects of gender and family history of alcohol dependence on a behavioral task of impulsivity in healthy subjects. J. Stud. Alcohol **63**(1), 83–90 (2002). https://doi.org/10.1007/s10964-021-01556-3
40. Engineering Toolbox: Required Space per Person (2003) (Accessed on 04.08.2022). https://www.engineeringtoolbox.com
41. Adli, M.: Urban stress and mental health. (2011) https://www.semanticscholar.org/paper/Urban-stress-and-mental-health/a78c67022ddbe961de35479f36c7438f9f70a35c. Accessed Nov 2011
42. Palma, A.d.: Optimal transit pricing with crowding and traffic congestion: a dynamic equilibrium analysis, Transportation Research Forum (TRF), 57th Transportation Research Forum (51st CTRF) Joint Conference, Toronto, Ontario, May 1–4. (2016) https://doi.org/10.22004/ag.econ.319276
43. Sharma, M. K. and Chauhan, B. K.: Analytical assessment of implementation aspect of regional rapid transit system. Decis. Anal. J. **4** (2022). https://doi.org/10.1016/j.dajour.2022.100093
44. Report of the sub-committee on operations and maintenance systems for metro railways, 2013. Submitted to Ministry of Urban Development, Govt. of India. https://mohua.gov.in/upload/uploadfiles/files/Report%203%20Operations%20and%20Maintenance%20systems.pdf. Accessed Nov 2011

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Manish Kumar Sharma[1,2,3] · **Manvendra Singh**[4,5,6,7]

✉ Manish Kumar Sharma
manish.dtutech@gmail.com

Manvendra Singh
manvendra.singh@ncrtc.in

[1] National Capital Region Transport Corporation, Delhi, India

[2] Gurukul Kangri University, Haridwar, Uttarakhand, India

[3] Delhi Technological University, Delhi, India

[4] CISO, National Capital Region Transport Corporation, Delhi, India

[5] ZH College of Engineering & Technology, Aligarh Muslim University, Aligarh, India

[6] Symbiosis University, Pune, India

[7] BITS, Pilani, Pilani, India

Emergent Materials
https://doi.org/10.1007/s42247-024-00755-5

ORIGINAL ARTICLE

Up-Conversion Luminescence in $Sr_9Y_2W_4O_{24}$:Ho^{3+}, Yb^{3+} Phosphors for applications in Temperature Sensing and w-LEDs

Sheetal Kumari[1] · Aman Prasad[2] · A. S. Rao[1] · R. K. Sinha[1]

Received: 25 April 2024 / Accepted: 26 May 2024
© Qatar University and Springer Nature Switzerland AG 2024

Abstract
Ho^{3+}-Yb^{3+} activated $Sr_9Y_2W_4O_{24}$ (SYW) phosphors were synthesized using a traditional solid-state reaction method. X-ray diffraction (XRD), scanning electron microscopy (SEM), and Fourier transform infrared spectroscopy (FT-IR) investigations were conducted to assess the phase, morphology and bond structure in the as-synthesised samples. To determine the optical bandgap values, diffuse reflectance spectra (DRS) for the as-titled phosphors have been obtained. Preliminary photoluminescent studies on the as prepared phosphors focus on UV–visible excitation and emission spectra, as well as observations of decay lifetimes. Secondly, the up-conversion (UC) emission spectra of as-titled phosphors, excited by a 980 nm laser were also recorded. The UC findings show that there are three bands at 488 nm and 551 nm, and 668 nm, corresponding to transitions of Ho^{3+} ions from $^5F_3$ to $^5I_8$, $^5F_4/^5S_2$ to $^5I_8$, and $^5F_5$ to $^5I_8$, respectively. UC luminescence is entirely a two-photon absorption process, as demonstrated by the fitting results of the curve exhibiting the relation between pump power and UC luminescence intensity. Furthermore, the emission spectra data yielded CIE coordinates that are located in the green region. The high thermal activation energy of the SYW phosphor substantially improves its thermal stability, as indicated by temperature dependent PL (TD-PL) spectra. Furthermore, the fluorescence intensity ratio (FIR) was employed to explore the optical sensing properties. Through the research, the bidirectional energy transmission in the UC mechanism was better understood, potentially leading to applications in solid state lighting and temperature sensors and UC applications of this phosphor.

Keywords XRD · Phosphor · Photoluminescence · Up-conversion · CIE coordinates, FIR

1 Introduction

Rare earth (RE) ions doped phosphor materials have seen a lot of usage recently because of their distinctive photoluminescence characteristics. These materials exhibit promising prospects in many domains, including fluorescent lamps, solid state lasers, light emitting diodes (LED), cathode ray tubes, optical thermometry, drug delivery, and bio-imaging [1–5]. The lanthanide ions exhibit rich multi-modal photoluminescence, including up-conversion (UC), downshifting (DS), contingent upon the various excitation radiations employed. UC refers to nonlinear optical processes that progressively absorb two or more photons and transform them into light emissions at a wavelength shorter than the excitation wavelength [1, 2, 6]. The UC phosphors often employ RE ions as their luminescence centres due to their abundance of electronic states, extended duration of intermediate meta-stable states, and energy levels stacked in a ladder-like manner [7]. However, in some specific applications, such as optical thermometry, the plentiful 4f electronic states of dopant ions typically result in multicolor emission with many emission peaks from f-f transitions. The behaviour of optical temperature sensors based on UC emission has been extensively researched in recent years due to its excellent sensitivity and accuracy. Optical thermometry is a fundamental term that refers to some lanthanide ions with thermally coupled energy levels (TCELs) [1, 8, 9]. The fluorescence intensity ratio (FIR) approach, that involves two thermally connected levels, is currently the most widely used mechanism for optical temperature detection. The thermally connected amounts of RE ions have narrow energy gap (100-2000 cm^{-1}), which varies depending on the lanthanide

✉ A. S. Rao
drsrallam@gmail.com

[1] Department of Applied Physics, Delhi Technological University, Bawana Road, Delhi 110042, India

[2] Department of Physics and Computer Science, Dayalbagh Educational Institute (DEI), Deemed University, Dayalbagh Agra-282005, India

Published online: 01 June 2024

ions [10]. Vega et al. investigated the near-infrared to visible up-conversion and second harmonic production in the $BaTiO_3$ phosphor doped with Ho^{3+}/Yb^{3+}[11]. Lojpur et al. also reported temperature sensing characteristics for Ho^{3+}/Yb^{3+}:Y_2O_3 phosphor [12]. Furthermore, to the best of our knowledge, temperature sensing characteristics have not been explored in the SYW: Ho^{3+}/Yb^{3+} ions co-doped phosphors.

Meanwhile, down-conversion (DC) is the opposite of up-conversion (UC), and it can be utilized in numerous applications such as solid-state lighting (SSL), lasers, optoelectronic devices, and plasma devices when an appropriate host lattice is doped with RE ions [13–17]. Compared to traditional fluorescent and incandescent light sources, SSL devices offer a number of improved features, including a longer lifespan, lower production costs, and an environmentally friendly design [18–21]. Currently, the most widely used commercial w-LEDs are created by combining a blue LED semi-conductive chip with a YAG: Ce^{3+} phosphor that emits yellow light. Many problems, including low colour rendering index (CRI) and high correlated colour temperature (CCT) values, are caused by the absence of red colour component [22, 23]. Phosphors that emit green, red, and blue (RGB) light in n-UV chips were created as a feasible method that could be applied for production of w-LEDs. This method is not widely used, though, as the commercial red phosphor (Y_2O_2S: Eu^{3+}) used in it is not very good at absorbing light from the n-UV semiconductor chip.

In material science, doping is a commonly used technical procedure that includes adding atoms to obtain hybrid materials with desired features and capabilities. Since rare earth (RE) ions usually generate multicoloured emissions in UV to IR regions due to numerous ladder-like energy levels, long-lived excited states, and outstanding green emission, they are typically thought of as suitable options for lighting and display fields[24–27]. However, due to its limited absorption, the singly doped Ho^{3+} UC phosphor has very low pump efficacy and emission intensity under NIR stimulation. Since Yb^{3+} ions have a greater absorption cross-section at around 980 nm and can effectively transmit the excitation energy to Ho^{3+} ions, they can serve as excellent sensitizers for Ho^{3+} ions [28–31]. As a result, combining Ho^{3+} with Yb^{3+} may significantly increase Ho^{3+}'s emission efficacy. The distance between the acceptor and donor ions will often vary with different doping concentrations. Thus, it is very important to maximise the Ho^{3+}/ Yb^{3+} doping concentration and boost energy transfer efficiency.

For the creation of promising materials, the choice of an appropriate lattice is significant. A wide range of host materials have been thoroughly investigated, such as phosphates, titanates, tungstates, aluminates, and molybdates etc. Tungstates are a considerable class of host material by virtue of the numerous benefits they offer, particularly their simplicity of preparation, low cost, remarkable chemical strengths, and excellent optical characteristics [32–35]. When exposed to UV light, tungstate, a type of self-activated phosphor, emits a wide range of hues. The low phonon energy of tungstate helps to prevent highly competitive non-radiative decay for dopants and improves luminosity, making it a good host for phosphors. Thus, the LBGW:0.12Sm^{3+} was revealed by Pei et al. to serve as an orange-red phosphor in an LED field [36]. Kamal et al. conducted study on the sample (BWO: 0.07Tb^{3+}) for latent fingerprint identification and anti-counterfeiting applications [37]. Zhao et al. investigated $LiCaYb(WO_4)_3$:Er^{3+}/Tm^{3+} with very sensitive thermometry for thermometer applications [38]. On the other hand, SYW: Ho^{3+}/ Yb^{3+} phosphor has not yet been thoroughly studied for DC & UC mechanism.

In the present work, we utilised an intuitive high-temperature solid-state technique to synthesis a range of Ho^{3+}/ Yb^{3+} activated SYW phosphors. Using XRD, SEM & UV–vis-NIR techniques, we have examined the structural, morphology & optical properties of the as synthesised phosphors. To examine the many vibrational bands that are present in the lattice, FT-IR investigations were carried out. PL spectra measurements were used to investigate the luminous characteristics of the co-doped SYW phosphors. For excitation wavelengths at 454 nm and 980 nm, the DC & UC spectra of the samples were captured. The stability of intense high purity green UC emission was investigated at various pumping powers and dopant concentrations. We provide a detailed illustration of the potential energy transfer mechanism of the co-doped Ho^{3+}/ Yb^{3+} phosphors. Temperature dependence of the UC spectra under 980 nm excitation was recorded and the activation energy of the host material was estimated thereby providing an idea about the thermal stability of the phosphor. FIR calculations were used to obtain maximum relative sensitivity (S_R) and absolute sensitivity (S_A) values. These aforementioned UC PL features of Ho^{3+}/ Yb^{3+} co-doped SYW phosphors have not been reported yet.

2 Synthesis and instrumentations

The phosphor samples for this study were synthesised using the high temperature solid-state reaction technique. A range of phosphors have been produced, including SYW: x Ho^{3+} (x = 1.0, 3.0, 5.0, 7.0, & 9.0 mol%) and SYW: x Ho^{3+}/ y Yb^{3+} (x = 7.0 mol% & y = 1.0, 3.0, 5.0, & 7.0 mol%). For the phosphor synthesis, the high-purity raw ingredients $SrCO_3$ (98%), Y_2O_3 (98%), WO_3 (99.9%), Ho_2O_3 (99.9%), and Yb_2O_3 (99.9%) were utilised. The raw materials for each composition were measured out using an electric balance, then combined with a pestle and agate mortar for one hour to achieve a homogenous combination. The sample was then well mixed and heated for

4 h at 1300 °C, 5 °C per minute, in an alumina crucible. After the mixture had naturally cooled to ambient temperature, the resultant powder was ground for a second time to conduct appropriate characterizations.

Using a Bruker D-8 Model Advance diffractometer that comes with a nickel-based filter and Cu-Kα as source, XRD was utilised to examine the crystal phase of the Ho^{3+}/ Yb^{3+} co-doped samples. The PerkinElmer Spectrum 2 equipment was utilised for FT-IR spectroscopy. The JEOL 7610 F Plus microscope was used to capture SEM and EDX images. With a resolution of 2.0 nm, a Jasco V-770 spectrophotometer was used for diffuse reflectance spectroscopy (DRS). To assess the DC spectra, Xenon lamp was employed as the excitation source having resolution of 1.0 nm along with a Jasco 8300 FP spectrofluorophotometer. Using a 980 nm laser, Horiba PTI Quanta Master has been utilised for up-conversion PL. A 150 MHz digital oscilloscope (HM 1507 Hameg Instruments) was used to measure the PL decay.

3 Result and discussion

3.1 XRD analysis

XRD patterns were utilised to identify the phase formation of the samples, and the results obtained from all of the as-prepared samples were identical. Figure 1(a) shows a series of XRD patterns of SYW: Ho^{3+}/ Yb^{3+} phosphors with varying doping concentrations in the diffraction 2θ range 15°-60° as well as the JCPDS Card No. 49–0568. The presence of broad, strong peaks indicates that the phosphor particles are well-crystallized after being synthesised. There were no peaks from impurities, and all of the diffraction peaks showed good coincidence with the JCPDS data. This result shows that doping Ho^{3+} and Yb^{3+} ions in SYW phosphors has been accomplished.

Using the most prevalent diffraction peak and the provided Debye Scherrer formula, the crystallite size has been calculated [39].

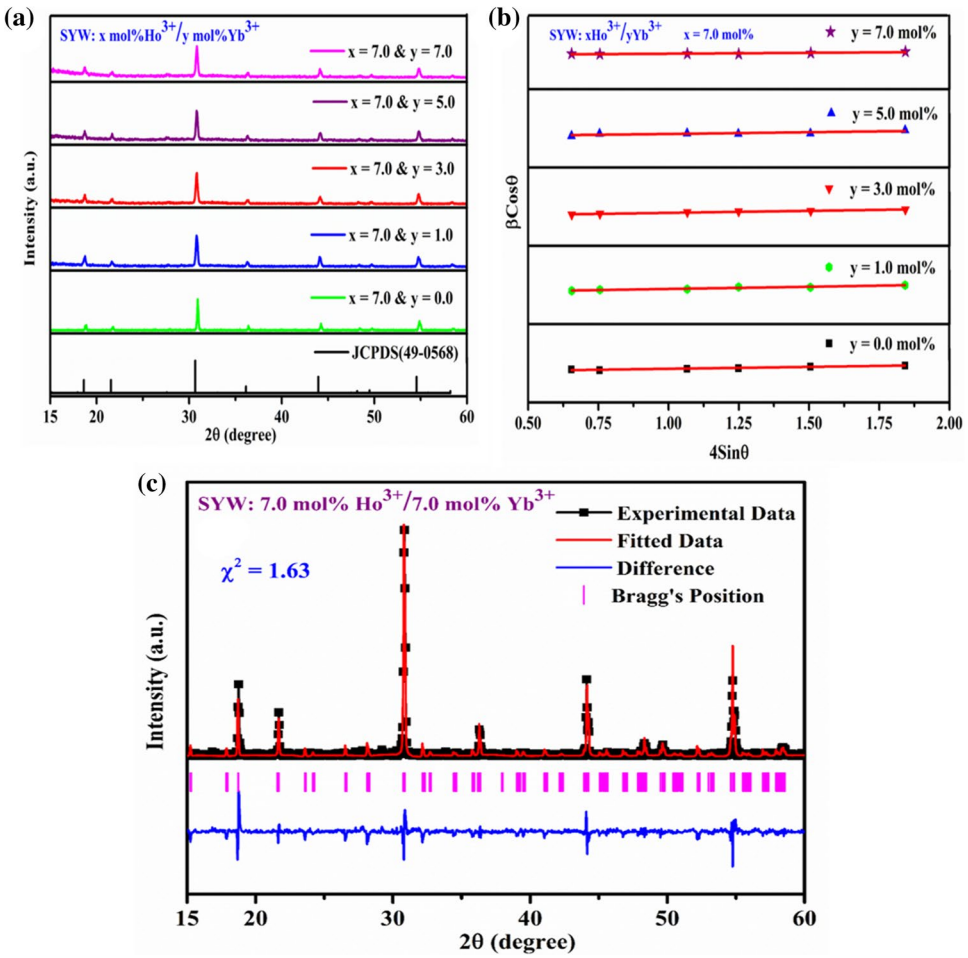


Fig. 1 (**a**)XRD patterns of SYW: xHo^{3+}/ yYb^{3+} (x = 7.0 mol% and y = 1.0, 3.0, 5.0 and 7.0 mol%) phosphors along standard JCPDS card. (**b**) their respective W–H plot of SYW phosphors & (**c**) show reitveld refinement recorded for SYW: 7.0 mol% Ho^{3+}/ 7.0 mol% Yb^{3+}) phosphors respectively

$$D = \frac{k\lambda}{\beta cos\theta} \quad (1)$$

where λ is the wavelength of the X-ray utilised, β indicates the full width at half maximum, and θ represents the diffraction angle, D is the crystallite size, and K is the form factor (0.9). The estimated crystallite sizes range from 22 to 27 nm.

It has been noted that the lattice strain (ε) and crystallite size may be affected by the β, (FWHM) of the XRD diffraction peak. This can be readily explained by the Williamson–Hall equation [39, 40],

$$\beta cos\theta = \frac{k\lambda}{D} + 4\varepsilon sin\theta \quad (2)$$

W–H plots are depicted in Fig. 1 (b). W–H plots and Debye Scherrer's formula have both been used to determine the crystallite size (D) and strain effect on SYW: Ho^{3+}, Yb^{3+} samples. Scherrer's analysis and W–H plots indicate that there isn't much fluctuation in crystallite size. This is due to the common assumption that the strain factor in Scherrer's equations is zero. The estimated strain in the as-synthesised phosphors in this instance is extremely small which has little influence on the doped samples' PL-intensity.

The Rietveld refinement analysis of SYW:x Ho^{3+},y Yb^{3+} (x = 7.0 mol% and y = 7.0 mol%) phosphor was conducted via the FULLPROF suit programme. The pseudo-Viogt function has been applied to fit different data point settings. The Rietveld refinement SYW:x Ho^{3+},y Yb^{3+} (x = 7.0 mol% and y = 7.0 mol%) is shown in Fig. 1(c). The calculated XRD pattern is shown by the solid line in the refinement diagram; the observed XRD patterns are represented by dots; the Bragg's positions of the cubic phase are represented by vertical bars; and the variation between the calculated and observed XRD patterns is displayed in the lower profile. Lower values for goodness of fit (GOF) = $\chi 2$ for the best refining results predicted the quality of the refine data Table 1.

Table 1 Rietveld refinement parameters with JCPDS data for SYW: 7.0 mol% Ho^{3+}/ 7.0 mol% Yb^{3+}) ion doped phosphor

Parameters	Standard data JCPDS (49–0568)	SYW phosphor
a(Å)	11.642	11.615
b(Å)	11.642	11.615
c(Å)	16.465	16.489
α = β = γ (°)	90	90
χ^2	–	1.63
Cell Volume (Å^3)	2231.60	2224.50

3.2 FESEM and EDX

To explore the atomic structure and particle size of the as-titled samples in further detail, FESEM has been employed. SEM micrographs of the Ho^{3+} doped and Ho^{3+}/Yb^{3+} co-doped SYW phosphors are displayed in Fig. 2(a & b). The uneven size and irregular polyhedron forms of SYW particles are clearly visible. Moreover, some phosphor particles exhibit very minimal agglomeration as a result of the high-temperature reaction process and hand grinding.The SYW phosphor's particle sizes as estimated from ImageJ software lay in the range of 4-9 µm...

An EDX analysis is used to identify the constituent elements of various materials. The single Ho^{3+} ion doped sample's EDX spectrum is displayed in Fig. 2(c), whereas the co-doped Ho^{3+}/ Yb^{3+} sample's spectrum is displayed in Fig. 2(d). The elements Sr, Y, W, Ho were detected in the analysis of singly doped e Ho^{3+} ions sample, and the elements Sr, Y, W, Ho, and Yb were detected in the Ho^{3+}/ Yb^{3+} ions co-doped sample. As a result, all of the components that were utilised for producing these samples are confirmed by the EDX record. The elemental mapping of the Ho^{3+}/ Yb^{3+} co-doped phosphor is displayed in Fig. 3. Every element occurs in the matrix, and the doping ions are evenly spaced over the surface, as can be seen from the elemental mapping.

3.3 FT-IR spectral analysis:

The nature of bonding and their stretching and bending vibrations in the as-titled phosphor was investigated using FT-IR analysis. The attenuated total reflectance mode (ATR) spectra of un-doped, SYW: 7.0 mol% Ho^{3+} and co-doped SYW: 7.0 mol% Ho^{3+}/ 7.0 mol% Yb^{3+} phosphors are shown in Fig. 4. There are bands around 564, 791, 857, 1062, 1470, 2892, and 2974 cm^{-1} in the spectra. Bands in the 470–735 cm^{-1} range are assigned to metal–oxygen (Y–O) bond vibrations, and the peak at 791 cm^{-1} is ascribed to the W–O-W stretching vibration mode. The peaks at approximately 857 cm^{-1} is due to the W–O bond's asymmetric stretching vibration. An infrared band located between 1062 and 1470 cm^{-1} indicates the existence of Sr–O. The occurrence of a hydroxyl group may be indicated by the band at 2892 cm^{-1}, and the antisymmetric stretching of adsorbed water molecules can be seen at 2974 cm^{-1}. Numerous functional groups have been observed to be the same in both doped and un-doped as-prepared phosphors.

3.4 DRS analysis and band gap energy

An important parameter of phosphors used in w-LEDs is their optical band gap (Eg). Figure 5 displays the DRS of undoped, 7.0 mol% Ho^{3+}, and 7.0 mol% Ho^{3+}/ Yb^{3+} doped SYW phosphors from 200–1400 nm. The spectra reveal

جامعة قطر QATAR UNIVERSITY

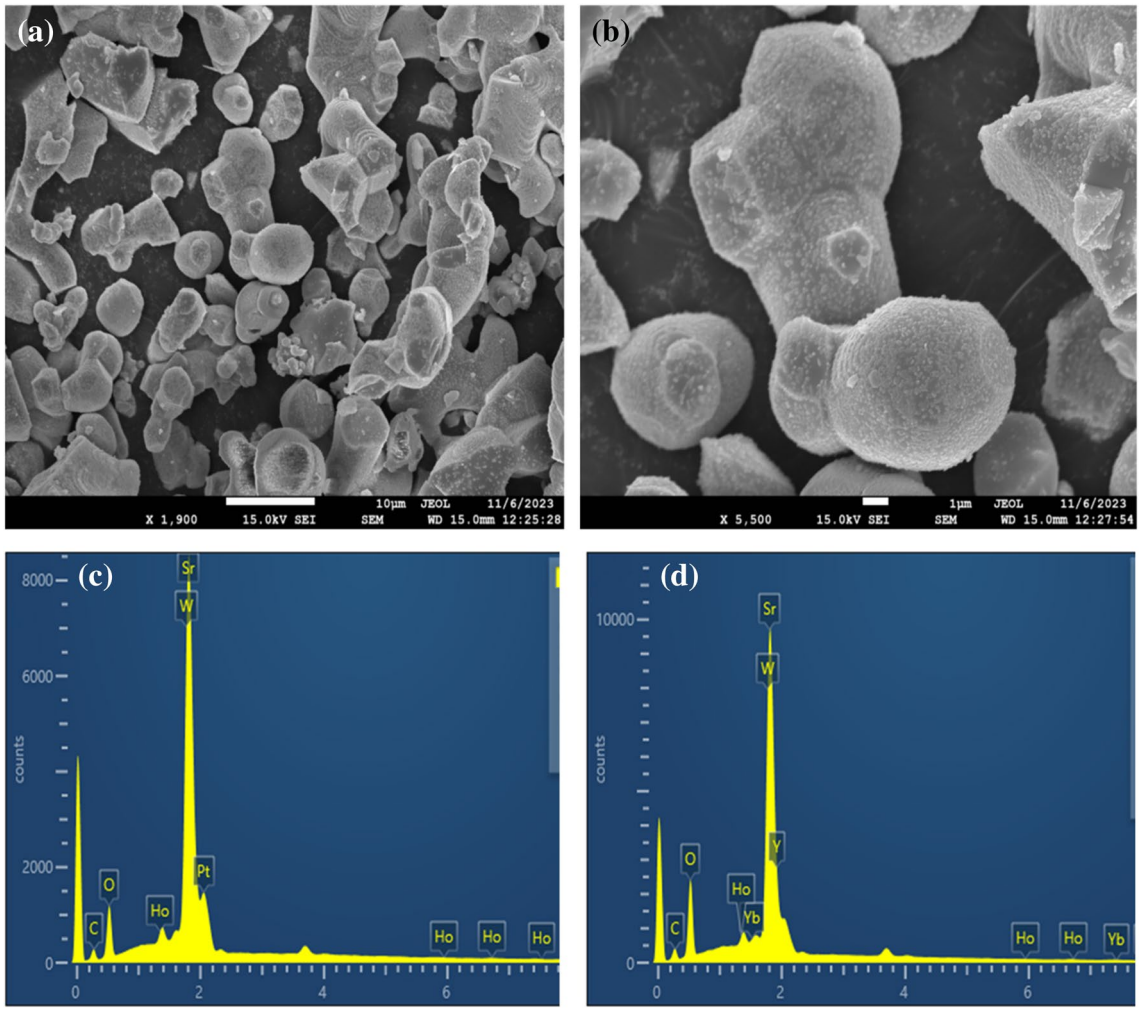


Fig. 2 SEM micrographs (**a**-**b**) of 7.0 mol% Ho^{3+} & 7.0 mol% Ho^{3+}/ 7.0 mol% Yb^{3+} co-doped SYW phosphors (**c**-**d**) and their corresponding EDX images

that the distinct, strong absorption bands at 454, 542, and 646 nm in the visible area are triggered by transitions from $^5I_8 \rightarrow {}^5G_6$, $^5I_8 \rightarrow {}^5F_4/{}^5S_2$, and $^5I_8 \rightarrow {}^5F_5$, in that order. In addition to this, the host also exhibits a prominent, broad absorption band in the UV area at 305 nm due to the charge transfer band (CTB) [41]. After converting the DRS spectral data into the Kubelka–Munk function, the optical band gap (E_g) was computed using the subsequent formulae [42, 43].

$$F(R) = \frac{(1-R^2)}{2R} = \frac{\alpha}{S} \tag{3}$$

The sample reflectance is denoted by R, the scattering coefficient by S, and the absorption coefficient by α.

$$F(R)h\nu = C(h\nu - E_g)^n \tag{4}$$

Here, h denotes Planck's constant, v represent the frequency of incident light, and C is an energy-independent constant (in this case, = 1). For direct & indirect allowed transitions, the value of n can be ½ and 2, respectively. The direct allowed optical bandgap (n = 1/2) for the current scenario has been computed. The graph $(\alpha h\upsilon)^2$ vs hυ of the undoped, Ho^{3+} doped, and Ho^{3+}/ Yb^{3+} co-doped phosphors is displayed as a Tauc plot in Fig. 6. By extrapolating the linear part of the plot, the optical band bap value for the samples was found to be between 4.36 and 4.52 eV. A lower band gap value leads to a greater number of ions being excited, which increases the intensity of photoluminescence.

3.5 PL properties

Figure 7(a) shows the PL excitation spectra of SYW: x Ho^{3+} (1.0, 3.0, 5.0, 7.0, & 9.0 mol %) phosphor, which was obtained by observing green emission at 551 nm. It has several bands that depict transitions from the ground state, $^5I_8$, to distinct excited states, $^3H_5$ (363 nm), $^5G_4$

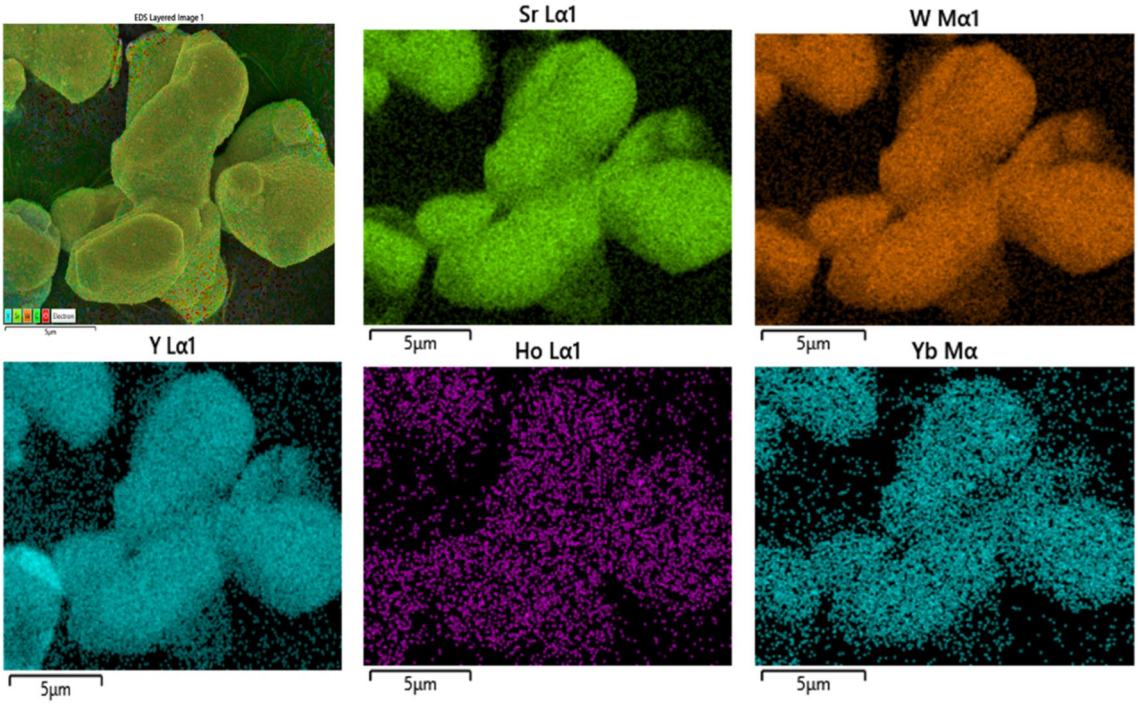


Fig. 3 Elemental mapping of 7 mol% Ho^{3+}/7 mol%Yb^{3+} co-doped SYW phosphor

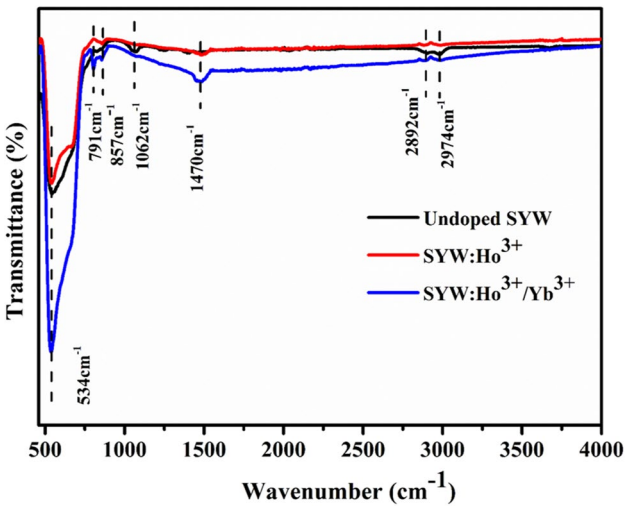


Fig. 4 FT-IR spectra of an un-doped, SYW: 7.0 mol% Ho^{3+} and co-doped SYW: 7.0 mol% Ho^{3+}/ 7.0 mol% Yb^{3+} phosphors

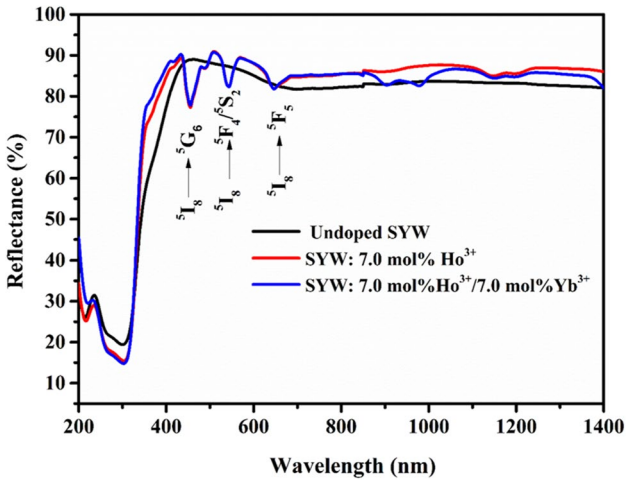


Fig. 5 DRS spectra for of the un-doped, 7.0 mol% Ho^{3+}, and 7.0 mol% Ho^{3+}/Yb^{3+} doped SYW phosphors

(387 nm), $^5G_5$ (421 nm), $^5G_6$ (454 nm), $^3K_8$ (476 nm), and $^5F_2$ (489 nm) [44, 45]. These transitions were noted in accordance with data reported by Carnall et al. [46, 47]. From these bands, the band at 454 nm ($^5I_8 \rightarrow {}^5G_6$) exhibited high intensity; thus, this transition wavelength was chosen to examine the visible photoluminescence (PL) spectra for varying concentrations of Ho^{3+} ion doped SYW phosphors. Figure 7(b) displays the PL emission spectra of Ho^{3+} doped SYW phosphors at different concentrations that were excited by blue light with a wavelength of 454 nm. The emission spectra shows emission band, exhibiting a maximum at 551 nm that is ascribed to the Ho^{3+} ion's $^5F_4/^5S_2 \rightarrow {}^5I_8$ transition [48–50]. Furthermore, the emission bands associated with the transition from excited states $^5F_5$ (667 nm) to ground state $^5I_8$ are much weaker than the bands assigned at 551 nm in the current emission spectra.

جامعة قطر QATAR UNIVERSITY

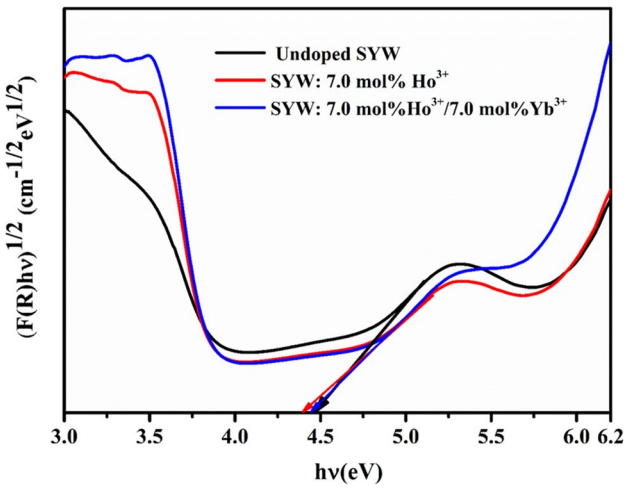


Fig. 6 The tauc's plot for direct band for SYW phosphors

The linear emission spectra obtained under 454 nm excitation for the SYW: x Ho^{3+}/y Yb^{3+} (x = 7.0 mol%, & y = 1.0, 3.0, 5.0 & 7.0 mol%) phosphors are displayed in Fig. 8. The co-doped SYW phosphors exhibit visible emission as well when excited with blue light. Three emission peaks are evident at wavelengths of 488, 551 and 668 nm, which corresponds to the transitions $^5F_3 \rightarrow {}^5I_8$, $^5F_4/{}^5S_2 \rightarrow {}^5I_8$, and $^5F_5 \rightarrow {}^5I_8$, respectively [24, 51]. As the Yb^{3+} concentration increases, it is evident from Fig. 8 that the emission intensity also increases. This could be because of the ET from Yb^{3+} to Ho^{3+} ions in the emission, and the cross relaxation between Ho^{3+} and Yb^{3+} ions. Furthermore the investigation on the UC behaviour of Ho^{3+}/ Yb^{3+} co-doped SYW phosphors carried out, since the Ho^{3+} ions exhibit very low or insignificant emission at excitation 980 nm.

Figure 9 demonstrates the UC spectra of the as-synthesised SYW: Ho^{3+}/ Yb^{3+} phosphors using 980 nm laser illumination in the 425–725 nm range. In the spectra, three distinct bands were seen in the blue, green, and red sections, at wavelengths of 488, 551, and 668 nm respectively. The subsequent transition of the Ho^{3+} ions from their $^5F_3$,$^5F_4/{}^5S_2$, and $^5F_5$ excited states to the $^5I_8$ ground state, is attributed by these blue, green, and red bands respectively [24, 52, 53]. The figure also clearly shows that as the Yb^{3+} concentration rises, so does the intensity of the UC emission. Ho^{3+} ions are known to have an extremely weak absorption cross-section when excited by a 980 nm laser. Consequently, Yb^{3+} ions are serving as a sensitizer in the current study and transmit energy to Ho^{3+} ions (the activator). Furthermore, it is evident from Fig. 9 that the UC emission produced by Ho^{3+} ions increases steadily up to 7 mol% when Yb^{3+} ion concentrations (1.0, 3.0, 5.0, and 7.0 mol%) rise. Doping Yb^{3+} ions causes energy to flow from Yb^{3+} ions to Ho^{3+} ions, which affects luminescent effectiveness. This is the cause of the increased red and green emission of Ho^{3+} ions. We have examined the power pump dependence PL studies to acquire additional confirmation for this up-conversion process.

3.6 Power pump dependent PL analysis

SYW: x Ho^{3+}/y Yb^{3+} (x = y = 7.0 mol%) was illuminated by variable power 980 nm laser to investigate the process of ET

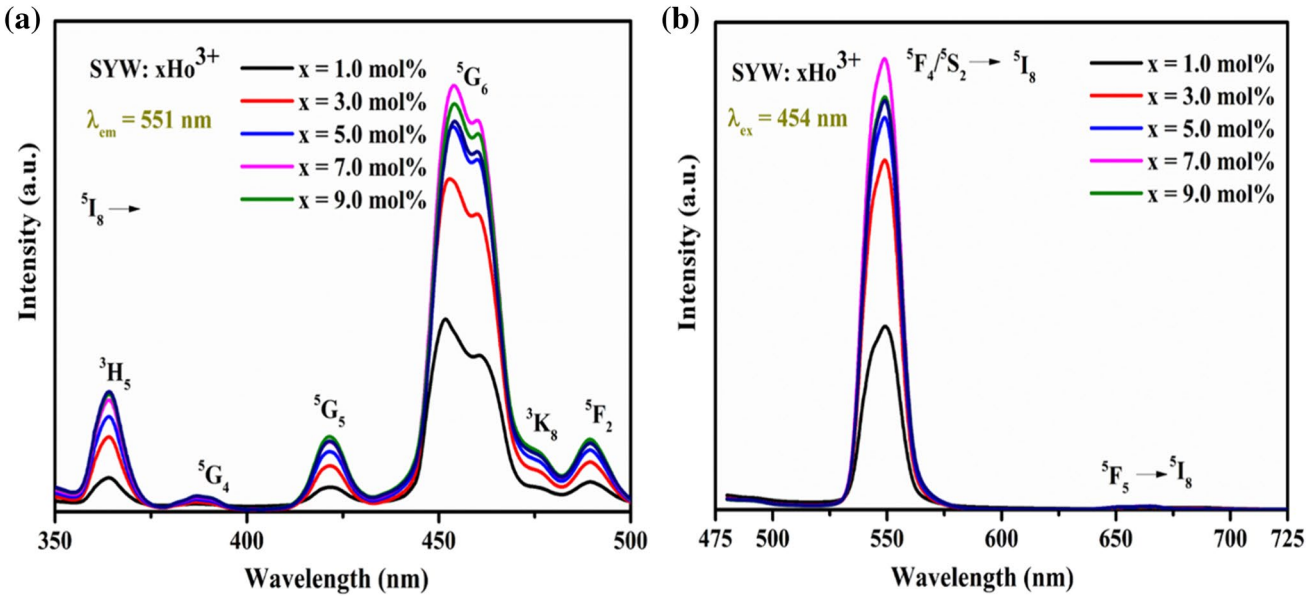


Fig. 7 (**a**) PL excitation spectra for SYW: xHo^{3+} (1.0, 3.0, 5.0, 7.0, & 9.0 mol %) doped phosphors at 551 nm wavelength. (**b**) PL emission spectra of SYW: xHo^{3+} (1.0, 3.0, 5.0, 7.0, & 9.0 mol %) doped phosphors under excitation wavelength 454 nm

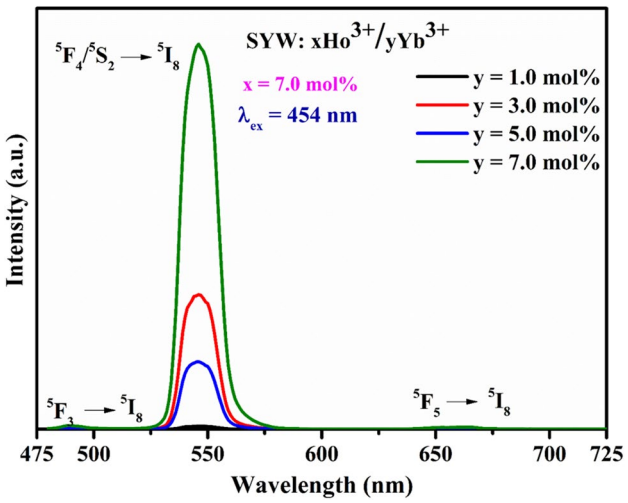


Fig. 8 Linear emission spectra under 454 nm excitation for the SYW: xHo^{3+}/yYb^{3+} (x = 7.0 mol%, & y = 1.0, 3.0, 5.0 & 7.0 mol%) phosphors

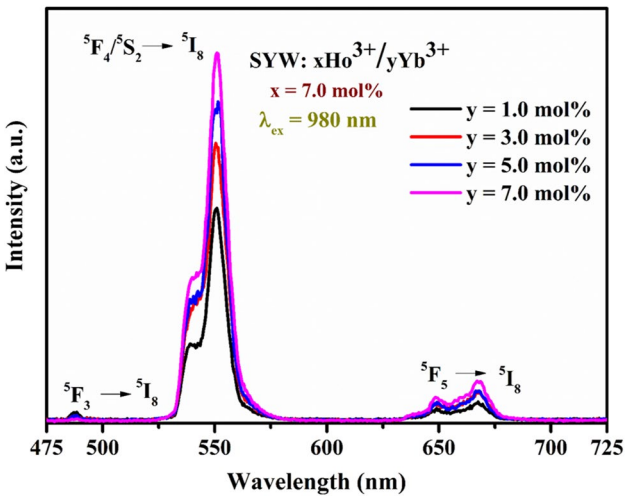


Fig. 9 UC spectra of the SYW: xHo^{3+}/yYb^{3+} (x = 7.0 mol%, & y = 1.0, 3.0, 5.0 & 7.0 mol%) phosphors under 980 nm excitation

between Yb^{3+} and Ho.$^{3+}$ ions in the sample. The variability in green and red band intensity with pump power is depicted in Fig. 10(a). In this up-conversion spectra, the green band is more intense than the red band. The UC intensity rises with increasing pump power at first, then saturates slightly. This intensity saturation can be the result of the material heating at more pump power. The emission intensity is directly related to nth power of pump power [54, 55]

$$I_{up} = (P_{in})^n \tag{5}$$

The variables I_{up}, P_{in}, and n represents the intensity of UC luminescence, pump laser, and number of needed pump photons, respectively. Figure 10(b) displays the outcomes of a logarithmic graph between power and intensity of the phosphor. For the emissions at 551 and 668 nm, the slope (n) values were 1.48 and 1.87, respectively. In the case of the two photon processes that are responsible for the observed green and red emissions, n ought to be almost equal to 2 which is confirmed from the plot. These findings suggest that the red and green emissions in phosphor are produced by two-photon processes.

3.7 Energy transfer mechanism

Using the aforementioned data as well as the Ho^{3+} and Yb^{3+} energy level diagrams, Fig. 11 illustrates potential mechanisms and population processes in the SYW: Ho^{3+}/ Yb^{3+} phosphor. When an infrared photon is absorbed by Yb^{3+} under the excitation of a 980 nm laser diode, electrons in the ground state of $^2F_{7/2}$ jump to the excited state of $^2F_{5/2}$. Ho^{3+} ions go through ground state excitation (GSA), which involves lifting an electron from the $^5I_8$ level to the $^5I_5$ level[56]. An additional lift of the electron from the $^5I_5$ level to the $^5F_2$ level occurs. The excited state absorption (ESA) of the $^5I_6$ level to the $^5S_2/^5F_4$ levels is also caused by the ET from Yb^{3+} to Ho^{3+} ions. Their population is increased by a non-radiative (NR) decay from the $^5I_5$ levels to the $^5I_6$ and $^5I_7$ levels, which also initiates the ESA from these levels to the $^5F_4$ and $^5F_5$ levels, respectively. Energy transfer (ET) from Yb^{3+} ions boosts each of these ESA even more. Green photon emission is produced when the highly populated $^5F_2$ level decays non-radiatively to the $^5F_4$ level, which decays radiatively[57]. Although the green emission predominates, the $^5F_2$ level also exhibits some radiative decay leading to an NIR photon emission. The non-radiative decay of $^5F_4$ level contributed slightly to the $^5F_5$ level, which is mostly populated by the ESA of $^5I_7$ level. As a result, there are fewer atoms at the $^5F_5$ level, and it releases a red photon as it relaxes to the ground state[51]. Conversely, the densely occupied $^5F_4$ level would rather relax to the ground level $^5I_8$ than to the $^5I_7$ level and release a green photon. As a result, the entire UC process is dominated by the emission of green photons.

3.8 CIE coordinates estimation

The CIE coordinates for the SYW: Ho^{3+}/ Yb^{3+} doped phosphors has been calculated using PL data from 1931CIE. The CIE diagram for SYW: Ho^{3+}/ Yb^{3+} doped phosphors with excitation wavelengths of 454 nm and 980 nm are displayed in Fig. 12(a) and (b), and Table 2 presents each of their estimated values. According to the CIE diagram, the coordinates of the as-synthesised phosphors are located close to the green region. Therefore, SYW: Ho^{3+}, Yb^{3+} could be a

جامعة قطر QATAR UNIVERSITY

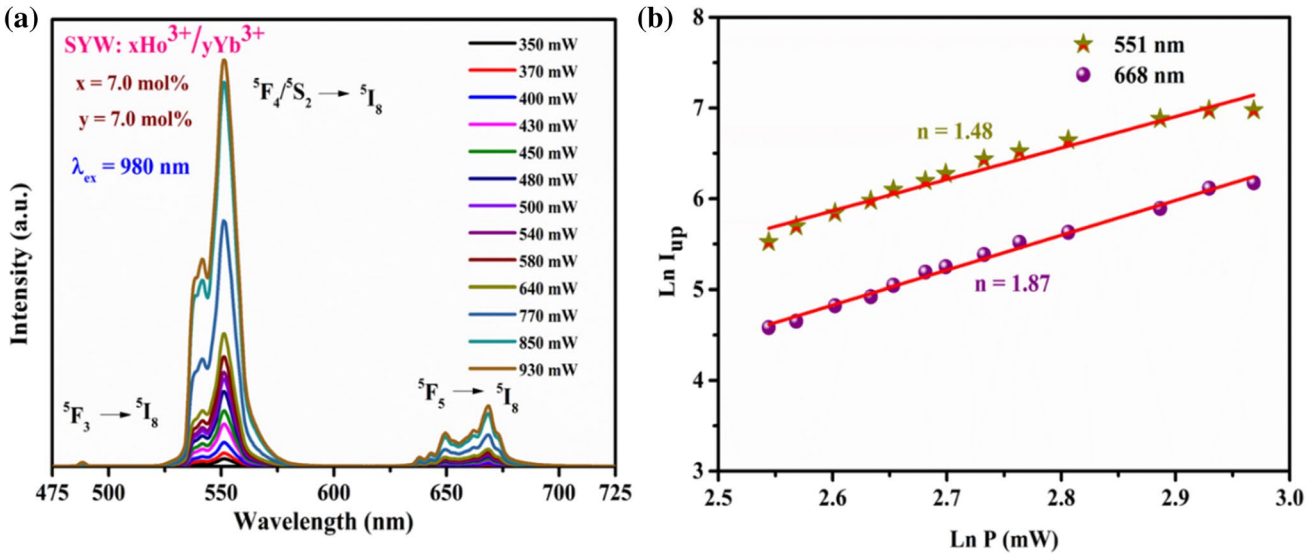


Fig. 10 (**a**)Power dependence PL spectra of SYW: xHo^{3+}/yYb^{3+} (x = 7.0 mol%, & y = 7.0 mol%) phosphor under 980 nm laser excitation and (**b**) plot of Ln I_{up} & Ln P

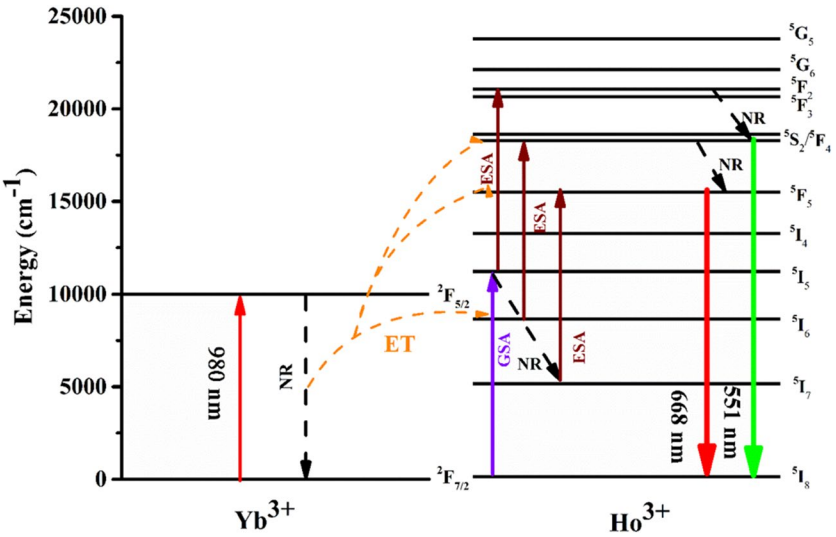


Fig. 11 Energy level diagram for Ho^{3+} & Yb^{3+} ions

good option with promise for lighting and multicolor display applications.

Typically, CCT is a useful metric for assessing the quality of a light source. The values of CCT can be calculated using McCamy's empirical formula, which is provided below [58, 59].

$$CCT = 449n^3 + 3525n^2 - 6823.2n + 5520.3 \quad (6)$$

Here, and xe = 0.332, ye = 0.186. The CCT values of Ho^{3+}/ Yb^{3+} doped SYW phosphor at λex = 980 nm are tabulated in Table 2.

Furthermore, one of the crucial factors for the w-LEDs application is colour purity. For the phosphor samples, the colour purity was determined using the following relation [17, 60].

$$Colorpurity = \frac{\sqrt{(x - x_{ee})^2 + (y - y_{ee})^2}}{\sqrt{(x_d - x_{ee})^2 + (y_d - y_{ee})^2}} \quad (7)$$

In this case, (x, y) represents the CIE coordinates of the SYW phosphor; (x_{ee}, y_{ee}) stands for the CIE coordinates of the standard white colour; and (x_d, y_d) represents the CIE coordinates of the dominant wavelength. The value of (x_i,y_i) was (0.33, 0.33). Table 2 also displays the computed colour purity values for the samples. It is noted that the Ho^{3+}, Yb^{3+} doped SYW samples colour purity lies between 97.4

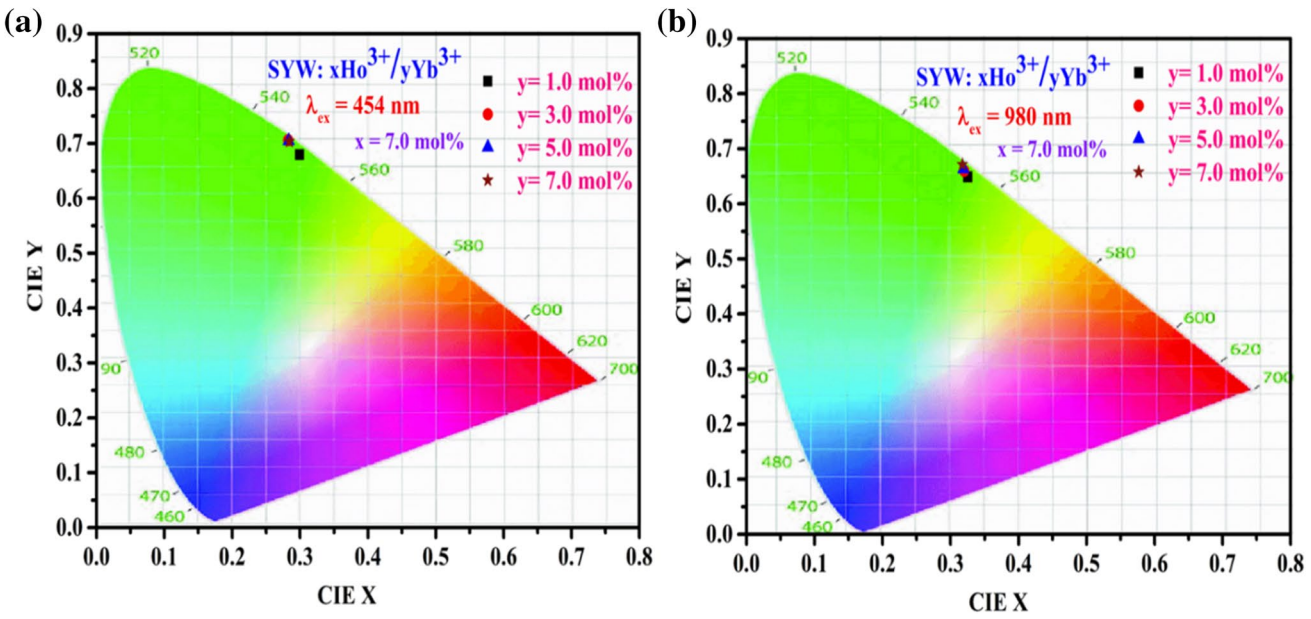


Fig. 12 CIE diagram for SYW:Ho^{3+}/Yb^{3+} doped phosphors with excitation wavelengths of 454 nm and 980 nm

Table 2 CIE coordinates, CCT values & color purity of SYW: xHo^{3+}/yYb^{3+} (x = 7.0 mol%, & y = 1.0, 3.0, 5.0 & 7.0 mol%) phosphors under 454 nm & 980 nm excitation

Sample ID SYW	CIE coordinates (x, y) λ_{ex} = 980 nm	CCT values (K) λ_{ex} = 980 nm	CIE coordinates (x, y) λ_{ex} = 454 nm	CCT values (K) λ_{ex} = 454 nm	Color purity (%)
y = 1.0 mol%	(0.3382, 0.6520)	5433.13	(0.2993,0.6795)	5994.13	97.6
y = 3.0 mol%	(0.3412, 0.6508)	5390.01	(0.28268,0.7058)	6209.19	97.9
y = 5.0 mol%	(0.3477, 0.6453)	5294.23	(0.2835,0.7048)	6197.77	98.2
y = 7.0 mol%	(0.3503, 0.6427)	5254.67	(0.2842,0.7038)	6188.50	97.4

and 98.2%. High colour purity values strongly imply that the synthesised samples have a lot of potential for use in w-LED applications. Therefore, SYW: Ho^{3+}, Yb^{3+} could be a good option with promise for lighting and multicolor display applications.

3.9 Decay profiles

Figure 13 illustrates the fluorescence decay profiles of the SYW: x Ho^{3+}/y Yb^{3+} (x = 7.0 mol % and y = 1.0, 3.0, 5.0, & 7.0 mol%) phosphors in Ho^{3+}/ Yb^{3+} co-doped samples, which were recorded at 980 nm excitation. Many exponential equations were used to fit the decay curves; in the present scenario, the bi-exponential equation fit the data the best [61].

$$I = I_0 + A_1 * exp\left(-\frac{t}{\tau_1}\right) + A_2 * exp(-\frac{t}{\tau_2}) \quad (8)$$

where I_0 & I signify the luminescence intensities at time t = 0 and at some other time, respectively, τ_1 & τ_2 are the components for rapid and slow decay, A_1 and A_2 are the fit constants for bi-exponential fitting. The average lifetime values for the bi-exponential fit are given by [62]

$$\tau_{exp} = \frac{A_1\tau_1^2 + A_2\tau_2^2}{A_1\tau_1 + A_2\tau_2} \quad (9)$$

SYW: x Ho^{3+}/y Yb^{3+} phosphors have an effective average decay time of 69.89, 67.61, 53.88, and 41.23 µs for x = 7.0 mol% and y = 1.0, 3.0, 5.0, & 7.0 mol%, respectively. As the concentration of Yb^{3+} ions rises, the decay time of Ho^{3+} ions decreases progressively, implying that ET from sensitizer to activator ions. Because of the presence of an ET channel, the Yb^{3+} ions transferred their energy to Ho^{3+} ions in the phosphor lattice.

3.10 Temperature sensing behaviour

One of the most important aspects in the high power w-LEDs and temperature sensing is the phosphor's PL thermal stability. The TD-PL spectra of SYW: x Ho^{3+}/y

جامعة قطر QATAR UNIVERSITY

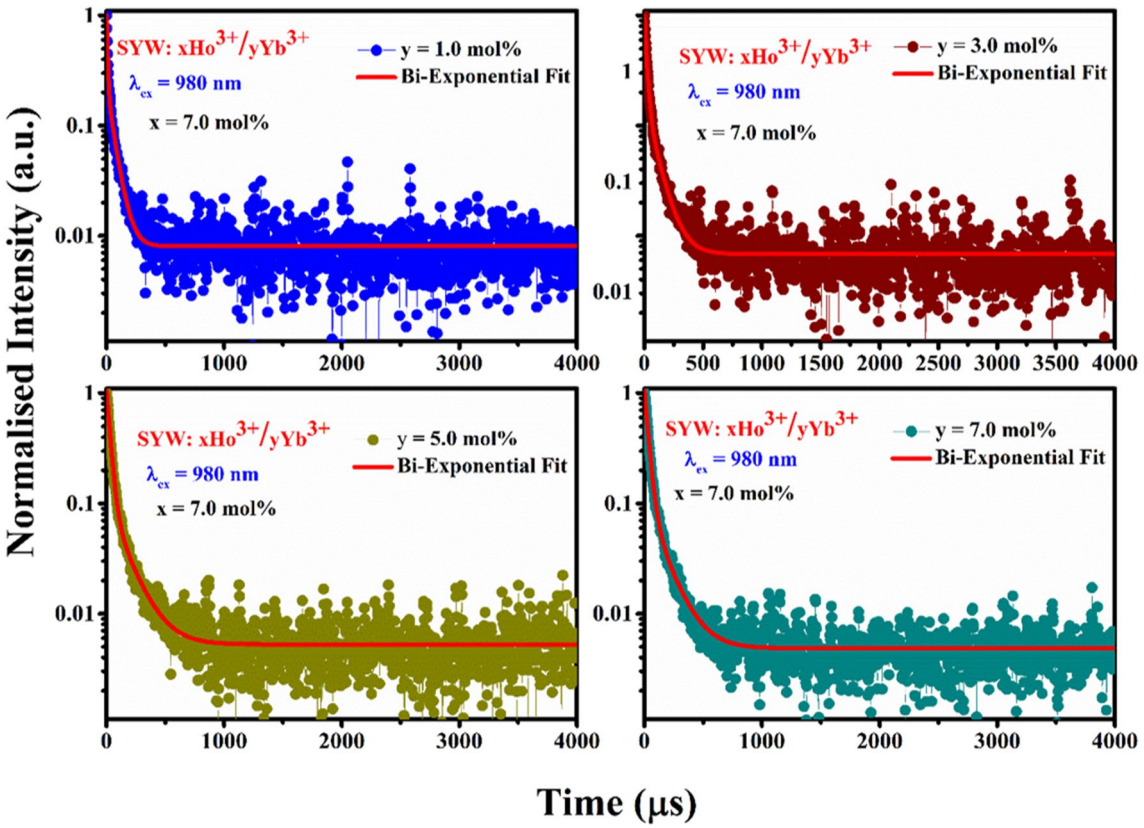


Fig. 13 The decay curves of SYW: xHo^{3+}/yYb^{3+} (x = 7.0 mol%, & y = 1.0, 3.0, 5.0 & 7.0 mol%) phosphors under 980 nm excitation

Yb^{3+} (x = y = 7.0 mol%) of phosphor was monitored under the excitation of 980 nm, in the range of 303 K to 453 K. The TD-PL emission intensity change with temperature for this phosphor is depicted in Fig. 14(a). According to the data, it seemed that when temperature increased, the emission intensity decreased. The phosphor was found to retain around 57% of its original intensity (303 K) at 453 K. Therefore, it is safe to assert that this phosphor is thermally stable and appropriate for use in white-LED applications.

The Arrhenius equation's exponential form can be used to estimate the activation energy (ΔE) for the thermal quenching process [63].

$$I_T = \frac{I_0}{1 + Cexp\left(-\frac{\Delta E}{K_B T}\right)} \tag{10}$$

where T is the temperature, c denotes the constant term, k represents the Boltzmann constant, and I_0 and I_T are the initial and emission intensity at different temperatures, respectively. Figure 14(b) displays a graph of ln [(I_0/I)—1] vs $1/k_BT$. Based on the plot, the activation energy, or ΔE, of SYW phosphor is 0.276 eV that is greater the other reported data [64–66]. The aforementioned findings indicate that the single phased phosphor can be used in w-LEDs and temperature sensing behaviour.

The population of the two thermally connected levels could be described using the Boltzmann distribution with temperature as a parameter. The intensity ratio of emissions from the two thermally connected levels can utilised to determine their respective populations. The FIR curve is formed by fitting the integral FIR of the two emission bands (at 551 nm and 668 nm), which can be used to determine how effectively the phosphor functions as a thermometer. Figure 15 depicts the fluctuation in integrated intensity as temperature increases.

Furthermore, the FIR curve is fitted using the equation shown below [67]:

$$FIR = \frac{I_0}{I} = Aexp(\frac{\Delta E_{TCEL}}{k_B T}) \tag{11}$$

Here, I_0 and I are the TD-PL intensities for the Ho^{3+}/Yb^{3+} ion transitions $^5F_4/^5S_2 \rightarrow {}^5I_8$ and $^5F_5 \rightarrow {}^5I_8$, respectively, A represents the fitting constant, ΔE is the splitting energy between TCELs, k_B stands for Boltzmann constant, and T denotes the temperature. Figure 16(a) depicts the I_{551}/I_{668} dependency on absolute temperature.

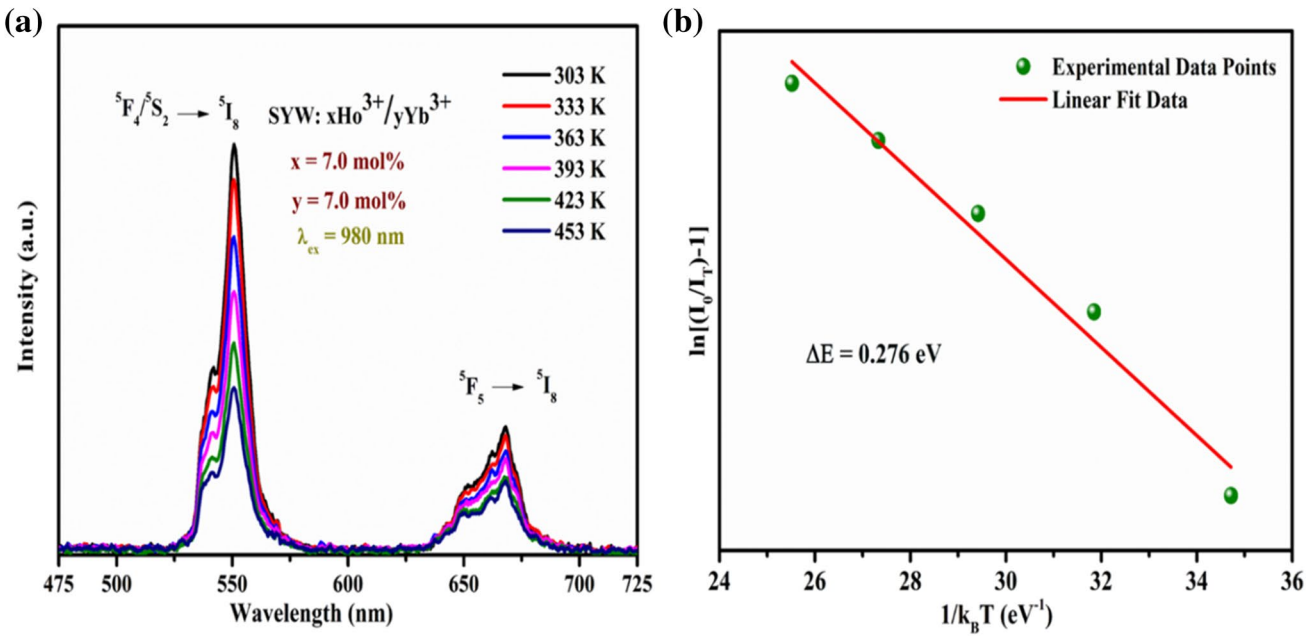


Fig. 14 (**a**) TD-PL spectra of SYW: 7.0 mol% Ho^{3+}/7.0 mol% Yb^{3+} phosphor using 980 nm laser source (**b**) Linear plot of ln $[(I_0/I_T)-1]$ vs $1/k_BT$ plot

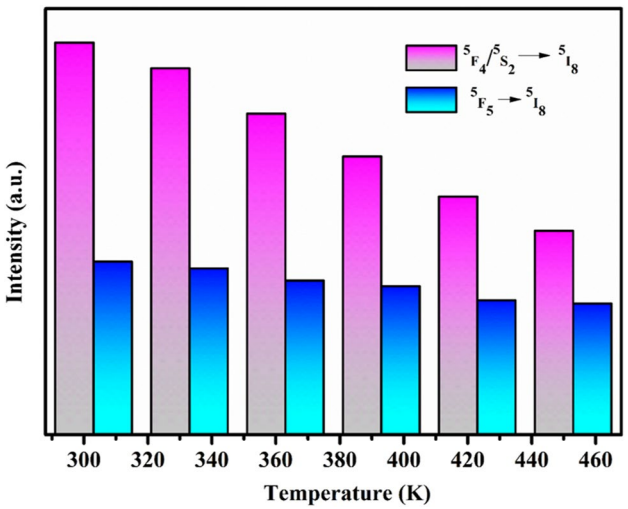


Fig. 15 Variation of the intensity of the emission peaks $^5F_4/^5S_2 \rightarrow {}^5I_8$ and $^5F_5 \rightarrow {}^5I_8$ with temperature

Fitting the experimental data can reveal a relation that is linear. The slope derived from the data points by linear fitting is 367.48. The energy difference between TCELs was calculated using the slope and k_B values and found to be 255.40 cm^{-1}. The estimated ΔE_{TCEL} values were almost completely related with the energy population in Ho^{3+}/Yb^{3+} ions.

The following equation represents temperature sensitivity, an important property in optical thermometric measurements that describes the rate of change in the FIR with temperature [68, 69].

$$S_A = \frac{dFIR}{dT} = FIR \times \frac{\Delta E_{TCEL}}{k_B T^2} \tag{12}$$

$$S_R = \frac{1}{FIR}\left(\frac{dFIR}{dT}\right) = \frac{\Delta E_{TCEL}}{k_B T^2} \tag{13}$$

Figure 16(b) depicts the variation of relative (S_R) and absolute (S_A) sensitivities with absolute temperature, with maximum values of 0.0091 K^{-1} and 0.0040 K^{-1}, respectively, at 303 K. Table 3 compares the optical temperature sensing capabilities of several UC hosts doped with Ho and Yb [10, 67, 68, 70]. The variation in sensitivity results could possibly be attributed to the diverse experimental circumstances of Ho^{3+}/Yb^{3+} UC lattices. Nevertheless, the as-titled phosphors used in this work can work as an effective optical temperature sensor as well.

4 Conclusion

In conclusion, a conventional high-temperature solid state reaction technique was utilised to successfully synthesise $Sr_9Y_2W_4O_{24}$: Ho^{3+} doped and $Sr_9Y_2W_4O_{24}$: Ho^{3+}/ Yb^{3+} co-doped phosphors. The formation of phase and crystal nature of the samples were verified using the XRD technique. The size of phosphor particles were found to be in the micrometre

جامعة قطر QATAR UNIVERSITY

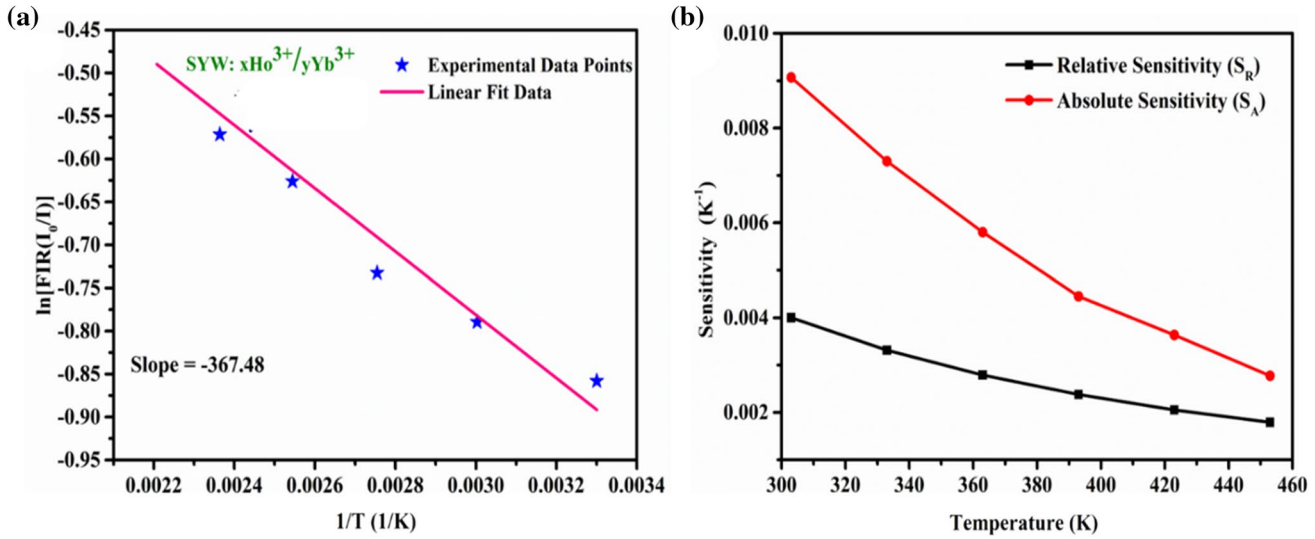


Fig. 16 (**a**) Plot of ln[FIR (I_0 /I)] vs 1/T (**b**) Relative sensitivity (S_R) & absolute sensitivity (S_A) values with temperature (K)

Table 3 Maximum S_R values of phosphor materials reported

Phosphor materials	Temperature Range (K)	S_R (% K^{-1})	References
$ZnGa_2O_4$:Ho^{3+}/Yb^{3+}/Bi^{3+}	303–600	0.0013	[10]
BaTiO3: Ho^{3+}/Yb^{3+}	305–515	0.0030	[65]
BYS: Ho^{3+}/Yb^{3+}	293–553	0.0049	[66]
NYSO: Ho^{3+}/Yb^{3+}	298–548	0.0153	[68]
SYW: Ho^{3+}/Yb^{3+}	303–453	0.0040	Present work

range, based on FESEM analysis and elemental mapping revealed that all of the essential elements required for synthesis were present in the lattice of the phosphors. Furthermore, it was been found that the components of the produced phosphor are evenly dispersed throughout its surface. The FT-IR spectra of the phosphors described the presence of different vibrational bonds and hydroxyl groups. The optical band gap values in the SYW host, determined by via Tauc plot, that turned out to be in the range of 4.36—4.52 eV. Under the excitation of 980 nm light, the green emission at 551 nm was seen, with a minimal red emission at 668 nm as well. This was due to the radiative transitions from $^5F_4$/ $^5S_2$ and $^5F_5$ excited states to the $^5I_8$ ground state, respectively. The amount of photons involved in the UC process can be obtained by analysing the variation in intensity and pump power. As the concentration of Yb^{3+} ions increased, the decay lifetime of Ho^{3+} ions at λ_{em} = 551 nm, which was monitored under 980 nm excitation, decreased. Every sample has CIE coordinates that are located in the green area under both 454 nm and 980 nm excitation. These samples also exhibit a high color purity. The TD-PL findings under 980 excitation demonstrates that the as-titled phosphor is relatively thermally stable having maximum relative sensitivity (S_R) and absolute sensitivity (S_A) values at 303 K temperature to be 0.0091 K^{-1} and 0.0040 K^{-1}, respectively. From the aforementioned results of the DC/UC studies conducted on the SYW phosphors, we wish to project these phosphors for usage in near-UV /IR stimulated white LEDs, temperature sensing and up-conversion applications.

Acknowledgements One of the authors, Ms. Sheetal Kumari is very much thankful to CSIR/UGC for the award of a Senior Research Fellowship (SRF) (UGC Ref. 1431/CSIR-UGC NET JUNE 2019). We are really grateful to Dr. Manju Srivastava and Dr. Shyama Prasad for the SEM facility provided by the Department of Chemistry at Dayalbagh Educational Institute (DEI), Deemed University, Agra.

Author contributions Ms. Sheetal Kumari: work plan, experiments, calculation and writing the manuscript.

Dr. Aman Prasad: review.

Prof. A.S. Rao: work plan, supervision, validation, correction of manuscript.

Prof. R. K. Sinha: work plan, supervision, validation, correction of manuscript.

Data availability The data that support the findings of this study are available on request from the corresponding author Prof. A.S. Rao. The data are not publicly available due to restrictions, e.g. their containing information that could compromise the privacy of research participants.

Declarations

Competing Interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. F. Huang, Y. Gao, J. Zhou, J. Xu, Y. Wang, Yb^{3+}/Er^{3+} co-doped $CaMoO_4$: A promising green upconversion phosphor for optical

temperature sensing. J. Alloys Compd. **639**, 325–329 (2015). https://doi.org/10.1016/j.jallcom.2015.02.228
2. B. Tio, X.L.I. Iaowei, C.H.J. Iang, Z.H.Z. Hai, X.L. V Uewei, Enhanced non-contact optical temperature sensing performance based on upconversion luminescence in Ho^{3+}/Yb^{3+} codoped ferroelectric $0.94Na_{0.5}Bi_{0.5}TiO_3$–$0.06BaTiO_3$, **13** 1626–1639 (2023). https://doi.org/10.1364/OME.489537
3. N. Wang, Z. Fu, Y. Wei, T. Sheng, Investigation for the upconversion luminescence and temperature sensing mechanism based on $BiPO_4$: Yb^{3+}, RE^{3+} (RE^{3+} = Ho^{3+}, Er^{3+} and Tm^{3+}). J. Alloys Compd. **772**, 371–380 (2019). https://doi.org/10.1016/j.jallcom.2018.09.070
4. B.N. Swathi, B.R.R. Krushna, S.A. Hariprasad, C. Srikanth, B. Subramanian, B.D. Prasad, H. Nagabhushana, Designing vivid green $Sr_9Al_6O_{18}$: Er^{3+} phosphor for information encryption and nUV excitable cool-white LED applications. J. Lumin. **257**, 119618 (2023). https://doi.org/10.1016/j.jlumin.2022.119618
5. B. Baby, S. Thomas, K. T, J. Jose, P.R. Biju, C. Joseph, Optimized microwave assisted synthesis of $La_2(WO_4)_3$:Tb^{3+} phosphors and analysis of photoluminescence behavior, J. Lumin. **254**, 119510 (2023). https://doi.org/10.1016/j.jlumin.2022.119510.
6. J. Song, T. Yang, H. Zhang, Y. Xiang, R. Song, B. Wang, J. Zhu, Thermally stable downconversion luminescence and upconversion temperature sensing properties of erbium-doped $Na_2Y_2TeB_2O_{10}$ phosphor. J. Lumin. **257**, 119664 (2023). https://doi.org/10.1016/j.jlumin.2022.119664
7. P. Liu, F. Wang, X. Chen, B. Yang, Upconversion/Downconversion luminescence and preparation of NIR-to-UV-excited Gd_2O_2S: Er phosphor. J. Lumin. **200**, 126–132 (2018). https://doi.org/10.1016/j.jlumin.2018.03.090
8. X. Wang, Y. Wang, L. Jin, Y. Bu, X.L. Yang, X. Yan, Controlling optical temperature detection of $Ca_3Al_2O_6$: Yb^{3+}, Er^{3+} phosphors through doping. J. Alloys Compd. **773**, 393–400 (2019). https://doi.org/10.1016/j.jallcom.2018.09.229
9. X. Cheng, X. Dong, K. Peng, H. Zhang, Y. Su, L. Jiang, Upconversion Luminescence and Optical Temperature-Sensing Properties of $LaNbO_4$: Yb^{3+}/ Er^{3+} Phosphors. J. Electron. Mater. **49**, 518–523 (2020). https://doi.org/10.1007/s11664-019-07776-5
10. R.S. Monika, A. Yadav, S.B. Rai, Rai, NIR light guided enhanced photoluminescence and temperature sensing in Ho^{3+}/ Yb^{3+}/Bi^{3+} co-doped $ZnGa_2O_4$ phosphor. Sci. Rep. **11**, 1–17 (2021). https://doi.org/10.1038/s41598-021-83644-9
11. M. Vega, I.R. Martin, J. Llanos, Near-infrared to visible upconversion and second harmonic generation in $BaTiO_3$: Ho^{3+} and $BaTiO_3$: Ho^{3+}/ Yb^{3+} phosphors. J. Alloys Compd. (2019). https://doi.org/10.1016/j.jallcom.2019.07.311
12. V. Lojpur, M. Nikolic, L. Mancic, O. Milosevic, M.D. Dramicanin, Y2O3: Yb, Tm and Y2O3: Yb, Ho powders for low-temperature thermometry based on up-conversion fluorescence. Ceram. Int. **39**, 1129–1134 (2013). https://doi.org/10.1016/j.ceramint.2012.07.036
13. P. Phogat, S.P. khatkar, V.B. Taxak, R.K. Malik, Sm^{3+} doped $Bi_4MgO_4(PO_4)_2$:crystal and optoelectronic investigation of the solution combustion derived bright orange emanating novel nanophosphor for SSLs, Mater. Chem. Phys. **276**, 125389 (2022). https://doi.org/10.1016/j.matchemphys.2021.125389.
14. R.G. Bora, C.B. Palan, N.S. Sawala, G.V. Korpe, S.K. Omanwar, Photoluminescence properties of $KSrPO_4$: Sm^{3+} phosphor for SSL applications. J. Mater. Sci. Mater. Electron. **29**, 541–545 (2018). https://doi.org/10.1007/s10854-017-7945-8
15. S. Kumari, Anu, A. Prasad, P. Rohilla, A.S. Rao, Prospective applications of thermally stable Dy^{3+} doped potassium zinc strontium borate (KZSB) glasses in w-LEDs, J. Mater. Sci. Mater. Electron. 34 (2023). https://doi.org/10.1007/s10854-023-10272-6.
16. S. Chand, R. Mehra, V. Chopra, Recent advancements in calcium based phosphate materials for luminescence applications. J. Lumin. **252**, 119383 (2022). https://doi.org/10.1016/j.jlumin.2022.119383
17. S. Kumari, A.S. Rao, R.K. Sinha, Investigations on photoluminescence and energy transfer studies of Sm^{3+} and Eu^{3+} ions doped $Sr_9Y_2W_4O_{24}$ red emitting phosphors with high color purity for w-LEDs. J. Mol. Struct. **1295**, 136507 (2024). https://doi.org/10.1016/j.molstruc.2023.136507
18. B. Han, J. Ren, P.P. Teng, J.B. Zhu, Y. Shen, Z.A. Li, X. Zhu, X. Yang, Synthesis and photoluminescence properties of a novel double perovskite Ca_2LaSbO_6: Sm^{3+} phosphor for w-LEDs. Ceram. Int. **48**, 971–980 (2022). https://doi.org/10.1016/j.ceramint.2021.09.182
19. K. Poria, R. Lohan, S. Bhatia, A. Kumar, R. Singh, N. Deopa, R. Punia, J.S. Shahi, A.S. Rao, Lumino-structural properties of Dy^{3+} activated $Na_3Ba_2LaNb_{10}O_{30}$ phosphors with enhanced internal quantum yield for w-LEDs. RSC Adv. **13**, 11557–11568 (2023). https://doi.org/10.1039/d3ra01260c
20. P. Rohilla, A.S. Rao, Synthesis optimisation and efficiency enhancement in Eu^{3+} doped barium molybdenum titanate phosphors for w-LED applications. Mater. Res. Bull. **150**, 111753 (2022). https://doi.org/10.1016/j.materresbull.2022.111753
21. J. Zhong, X. Chen, D. Chen, M. Liu, Y. Zhu, X. Li, Z. Ji, A novel rare-earth free red-emitting $Li_3Mg_2SbO_6$:Mn^{4+} phosphor-in-glass for warm w-LEDs: Synthesis, structure, and luminescence properties. J. Alloys Compd. **773**, 413–422 (2019). https://doi.org/10.1016/j.jallcom.2018.09.162
22. G. Wu, J. Xue, X. Li, Q. Bi, M. Sheng, Z. Leng, A novel red-emitting $Na_5W_3O_9F_5$: Eu^{3+} phosphor with high color purity for blue-based WLEDs. Ceram. Int. **49**, 10615–10624 (2023). https://doi.org/10.1016/j.ceramint.2022.11.249
23. S. Ye, F. Xiao, Y.X. Pan, Y.Y. Ma, Q.Y. Zhang, Phosphors in phosphor-converted white light-emitting diodes: Recent advances in materials, techniques and properties. Mater. Sci. Eng. R Reports. **71**, 1–34 (2010). https://doi.org/10.1016/j.mser.2010.07.001
24. G.B. Nair, A. Kumar, H.C. Swart, S.J. Dhoble, Facile precipitation synthesis of green-emitting BaY2F8: Yb3+, Ho3+ upconverting phosphor. Ceram. Int. **45**, 14205–14213 (2019). https://doi.org/10.1016/j.ceramint.2019.04.127
25. Y. Ato, R. Huzimura, L. Ozawa, Gamma-Ray Thermoluminescence in Rare-Earth-Activated Y_2O_3 Phosphors. Jpn. J. Appl. Phys. **7**, 1497 (1968). https://doi.org/10.1143/jjap.7.1497
26. A. Dwivedi, E. Rai, D. Kumar, S.B. Rai, E ff ect of Synthesis Techniques on the Optical Properties of Ho^{3+}/ Yb^{3+} Co-doped YVO_4 Phosphor : A Comparative Study, (2019). https://doi.org/10.1021/acsomega.8b03606.
27. V. Kumar, S. Som, S. Dutta, S. Das, H.C. Swart, Influence of Ho^{3+} doping on the temperature sensing behavior of Er^{3+}- Yb^{3+} doped La_2CaZnO_5 phosphor. RSC Adv. **6**, 84914–84925 (2016). https://doi.org/10.1039/c6ra13664h
28. H. Suo, C. Guo, W. Wang, T. Li, C. Duan, M. Yin, Mechanism and stability of spectrally pure green up-conversion emission in Yb^{3+}/ Ho^{3+} co-doped $Ba_5Gd_8Zn_4O_{21}$ phosphors. Dalt. Trans. **45**, 2629–2636 (2016). https://doi.org/10.1039/c5dt02985f
29. J. Tang, J. Gou, G. Li, Y. Li, H. He, C. Li, J. Yang, Synthesis, structure and upconversion luminescence of Yb^{3+}, Ho^{3+} co-doped $Gd_3Al_5O_{12}$ garnet phosphor prepared by the Pechini sol-gel method. RSC Adv. **6**, 54435–54439 (2016). https://doi.org/10.1039/c6ra09259d
30. J. Tang, C. Cheng, Y. Chen, Y. Huang, Yellow – green upconversion photoluminescence in Yb^{3+} ,Ho^{3+} co-doped $NaLa(MoO_4)_2$ phosphor, **609**, 268–273 (2014). https://doi.org/10.1016/j.jallcom.2014.04.134.
31. X. Wang, X. Li, S. Xu, L. Cheng, J. Sun, J. Zhang, B. Chen, Can temperature be accurately sensed by red-green emission ratio in $YNbO_4$: Ho^{3+}/ Yb^{3+} phosphor under 980 nm excitation? J. Alloys Compd. (2018). https://doi.org/10.1016/j.jallcom.2018.04.299

32. P. Du, L. Wang, J.S. Yu, Luminescence properties and energy transfer behavior of single-component $NaY(WO_4)_2$: Tm^{3+}/Dy^{3+}/Eu^{3+} phosphors for ultraviolet-excited white light-emitting diodes. J. Alloys Compd. **673**, 426–432 (2016). https://doi.org/10.1016/j.jallcom.2016.02.245
33. Q. Zheng, C. Guan, Z. Wan, R. Fu, J. Yang, Synthesis and up-conversion luminescence of Er^{3+} and Yb^{3+} co-doped $La_{14}W_8O_{45}$ phosphors, Opt. Mater. (Amst). 137 (2023). https://doi.org/10.1016/j.optmat.2023.113525.
34. F.B. Xiong, H. Chen, H.F. Lin, X.G. Meng, E. Ma, W.Z. Zhu, Photoluminescence characteristics of Sm^{3+} -doped $LnBWO_6$ (Ln = La, Gd and Y) as new orange-red phosphors. J. Lumin. **209**, 89–94 (2019). https://doi.org/10.1016/j.jlumin.2019.01.034
35. B. Han, B. Liu, Y. Dai, J. Zhang, Synthesis and luminescence properties of Sm^{3+} doped $NaKLaMO_5$ (M = Nb, Ta) red-emitting phosphors. Mater. Res. Bull. **121**, 110612 (2020). https://doi.org/10.1016/j.materresbull.2019.110612
36. X. Hong, P. Yan, Y. Zhao, K. Liu, Exploring a new Sm^{3+} - activated tungstate phosphor with thermally stable photoluminescence. J. Iran. Chem. Soc. **20**, 2269–2275 (2023). https://doi.org/10.1007/s13738-023-02832-8
37. R. Kamal, M. Saif, Barium tungstate doped with terbium ion green nanophosphor: Low temperature preparation, characterization and potential applications, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. **229**, 117928 (2020). https://doi.org/10.1016/j.saa.2019.117928.
38. Y. Zhao, X. Wang, Y. Zhang, Y. Li, X. Yao, Winning wide-temperature-range and high-sensitive thermometry by a multichannel strategy of dual-lanthanides in the new tungstate phosphors. J. Alloys Compd. **834**, 154998 (2020). https://doi.org/10.1016/j.jallcom.2020.154998
39. P. Kumar, S. Singh, I. Gupta, V. Kumar, D. Singh, Er^{3+} activated $LaAlO_3$ perovskite phosphor: Crystal structure and down conversion photoluminescent behaviour for optoelectronic devices. Inorg. Chem. Commun. **141**, 109578 (2022). https://doi.org/10.1016/j.inoche.2022.109578
40. P. Kumar, S. Singh, I. Gupta, K. Nehra, V. Kumar, D. Singh, Structural and luminescent behaviour of Dy(III) activated $Gd_3Al_5O_{12}$ nanophosphors for white-LEDs applications. Mater. Chem. Phys. **295**, 127035 (2023). https://doi.org/10.1016/j.matchemphys.2022.127035
41. Q. Wang, Z. Mu, D. Zhu, J. Feng, Z. Liang, L. Yang, X. Cai, J. Lu, F. Wu, A red emitting stoichiometric phosphor $Sr_2Ga_3La_{1-x}Eu_xGe_3O_{14}$ for white light emitting diodes. J. Mater. Sci. Mater. Electron. **28**, 15921–15927 (2017). https://doi.org/10.1007/s10854-017-7488-z
42. R.R. Cui, X. Guo, X.Y. Gong, C.Y. Deng, Photoluminescence properties and energy transfer of novel orange–red emitting phosphors: $Ba_3Bi_2(PO_4)_4$: Sm^{3+}, Eu^{3+} for white light-emitting diodes. Rare Met. **40**, 2882–2891 (2021). https://doi.org/10.1007/s12598-020-01663-3
43. Z. Liang, Z. Mu, Q. Wang, D. Zhu, F. Wu, The synthesis and luminescence properties of a novel red-emitting phosphor: Eu^{3+}-doped $Ca_9La(PO_4)_7$. Appl. Phys. A Mater. Sci. Process. **123**, 1–7 (2017). https://doi.org/10.1007/s00339-017-1226-5
44. V. Singh, K. V Dabre, S.J. Dhoble, G. Lakshminarayana, Optik Green emitting holmium (Ho) doped yttrium oxide (Y_2O_3) phosphor for solid state lighting, Opt. - Int. J. Light Electron Opt. **206,** 164339 (2020). https://doi.org/10.1016/j.ijleo.2020.164339.
45. P. Kaur, A. Kaur, S. Singh, L. Singh, Investigation on structural and thermoluminescence properties of Ho^{3+} doped SrB_4O_7 phosphor for dosimetry applications. J. Mol. Struct. **1248**, 131500 (2022). https://doi.org/10.1016/j.molstruc.2021.131500
46. W.T. Carnall, P.R. Fields, K. Rajnak, Electronic energy levels of the trivalent lanthanide aquo ions. IV. Eu $^{3+}$, J. Chem. Phys. **49**, 4424–4442 (1968). https://doi.org/10.1063/1.1669893.
47. W.T. Carnall, P.R. Fields, K. Rajnak, Electronic Energy Levels of the Trivalent Lanthanide Aquo Ions. III. Tb^{3+}, J. Chem. Phys. **49**, 4447–4449 (1968). https://doi.org/10.1063/1.1669895.
48. A. Marimuthu, R.N. Perumal, Investigations on structural, optical and electrical properties of holmium doped rubidium titanyl phosphate single crystals. J. Alloys Compd. **895**, 162568 (2022). https://doi.org/10.1016/j.jallcom.2021.162568
49. Y. Ye, Z. Tang, Z. Ji, H. Xiao, Y. Liu, Y. Qin, Fabrication and luminescent properties of holmium doped $Y_2Zr_2O_7$ transparent ceramics as new type laser material, Opt. Mater. (Amst). **121**, 111643 (2021). https://doi.org/10.1016/j.optmat.2021.111643.
50. Z. Mu, Y. Hu, G. Ju, Luminescence properties of Eu^{3+} and Ho^{3+} in Sr_2TiO_4. J. Rare Earths **30**, 744–747 (2012). https://doi.org/10.1016/S1002-0721(12)60122-7
51. K. Dharendra, A. Roy, A. Dwivedi, A. Kumar, S. Bahadur, D. Kumar, Two steps enhancement of dual mode (UC and DS) behaviour of Ho^{3+}/ Yb^{3+} and Tm^{3+}/ Yb^{3+} co-doped $GdVO_4$ phosphors : Improvement in spectral and color purity. Mater. Res. Bull. **139**, 111280 (2021). https://doi.org/10.1016/j.materresbull.2021.111280
52. P.K. Vishwakarma, A. Bahadur, A. Maurya, S.B. Rai, Large enhancement in upconverted green emission intensity from Ho^{3+}/ Yb^{3+} co-doped $Y_2Ti_2O_7$ phosphor in the presence of Zn^{2+}. Mater. Res. Bull. **115**, 219–226 (2019). https://doi.org/10.1016/j.materresbull.2019.03.029
53. P.K. Vishwakarma, S.B. Rai, A. Bahadur, Intense red and green emissions from Ho^{3+}/ Yb^{3+} co-doped Sodium Gadolinium Molybdate Nano-phosphor : Effect of calcination temperature and Intrinsic optical bistability. Mater. Res. Bull. **133**, 111041 (2021). https://doi.org/10.1016/j.materresbull.2020.111041
54. A. Dubey, A.K. Soni, A. Kumari, R. Dey, V.K. Rai, Enhanced green upconversion emission in NaYF4: Er3+/ Yb3+ /Li+ phosphors for optical thermometry. J. Alloys Compd. **693**, 194–200 (2017). https://doi.org/10.1016/j.jallcom.2016.09.154
55. C. Micale, R.M. Almeida, Structure and properties of Er and Er / Yb-doped YF_3 up-conversion phosphors compared with oxide hosts through an internal standard. Mater. Today Commun. **31**, 103239 (2022). https://doi.org/10.1016/j.mtcomm.2022.103239
56. Y. Zhou, H. Wang, B. Xu, L. Cao, Concentration-dependent up-converted luminescence of Ho^{3+}/ Yb^{3+} co-doped yttrium-stabilized zirconia phosphors. J. Lumin. **161**, 288–292 (2015). https://doi.org/10.1016/j.jlumin.2015.01.019
57. S. Liu, S. Liu, H. Ming, D. Hou, W. You, Investigation on the upconversion luminescence in Ho^{3+}/ Yb^{3+} co-doped $Ba_3Sc_4O_9$ phosphor. Mater. Res. Bull. **98**, 187–193 (2018). https://doi.org/10.1016/j.materresbull.2017.10.025
58. C.S. McCamy, Correlated color temperature as an explicit function of chromaticity coordinates. Color. Res. Appl. **17**, 142–144 (1992). https://doi.org/10.1002/col.5080170211
59. K.S. Rama, K. Reddy, K. Swapna, S. Mahamuda, M. Venkateswarlu, A.S. Rao, G.V. Prakash, Investigation on structural and luminescence features of Dy^{3+} ions doped alkaline-earth boro tellurite glasses for optoelectronic devices. Opt. Mater. (Amst) **85**, 200–210 (2018). https://doi.org/10.1016/j.optmat.2018.08.057
60. J. Wang, J. Guo, Q. Lv, Z. Ma, X. Feng, Y. Lu, J. Gao, W. Chen, B. Deng, R. Yu, Spectroscopic investigation of the novel orange-red phosphor Ca3La2W2O12: Sm3+ with the high color purity for w-LED applications. J. Lumin. **241**, 118459 (2022). https://doi.org/10.1016/j.jlumin.2021.118459
61. H.T. Haile, F.B. Dejene, Effect of substrate temperature on the material properties of the Y_2SiO_5: Ce^{3+} thin film by pulsed laser deposition (PLD)method. Optik (Stuttg). **184**, 508–517 (2019). https://doi.org/10.1016/j.ijleo.2019.05.003
62. J. Xie, L. Cheng, H. Tang, X. Yu, Z. Wang, X. Mi, Q. Liu, X. Zhang, Synthesis and photoluminescence properties of NUV-excited $NaBi(MoO_4)_2$: Sm^{3+} phosphors for white light emitting

diodes. Opt. Laser Technol. **147**, 107659 (2022). https://doi.org/10.1016/j.optlastec.2021.107659
63. S. Kumari, A.S. Rao, R.K. Sinha, Structural and photoluminescence properties of Sm^{3+} ions doped Strontium Yttrium Tungstate Phosphors for Reddish-orange Photonic device applications. Mater. Res. Bull. **167**, 112419 (2023). https://doi.org/10.1016/j.materresbull.2023.112419
64. S. Kumari, A.S. Rao, R.K. Sinha, Green Emission of Erbium Doped SYW Phosphors for Optical Thermometry And Solid-State Lighting. ChemPhotoChem. (2024). https://doi.org/10.1002/cptc.202300226
65. G. Zhu, Z. Li, F. Zhou, C. Wang, S. Xin, A novel temperature sensitive Sm^{3+} doped niobate orange-red phosphor : The synthesis and characteristic luminescent property investigation. J. Lumin. **196**, 32–35 (2018). https://doi.org/10.1016/j.jlumin.2017.12.014
66. L. Wu, Y. Bai, L. Wu, H. Yi, Y. Kong, Y. Zhang, J. Xu, Sm^{3+} and Eu^{3+} codoped $SrBi_2B_2O_7$: A red-emitting phosphor with improved thermal stability. RSC Adv. **7**, 1146–1153 (2017). https://doi.org/10.1039/c6ra26752a
67. P. Singh, R.S. Yadav, P. Singh, S.B. Rai, Upconversion and downshifting emissions of Ho^{3+}- Yb^{3+} co-doped $ATiO_3$ perovskite phosphors with temperature sensing properties in Ho^{3+}- Yb^{3+} co-doped $BaTiO_3$ phosphor. J. Alloys Compd. **855**, 157452 (2021). https://doi.org/10.1016/j.jallcom.2020.157452
68. J. Zhang, X. Li, G. Chen, Upconversion luminescence of $Ba_9Y_2Si_6O_{24}$: Yb^{3+}-Ln^{3+} (Ln = Er, Ho, and Tm) phosphors for temperature sensing. Mater. Chem. Phys. **206**, 40–47 (2018). https://doi.org/10.1016/j.matchemphys.2017.12.007
69. Z. Ni, Y. Zhu, H. Zhang, Y. Su, Yb^{3+}/ Ho^{3+} doped $Ba_3Y_2B_4O_{12}$ phosphors: Upconversion luminescence and temperature sensing properties. J. Lumin. **268**, 120427 (2023). https://doi.org/10.1016/j.jlumin.2023.120427
70. J. Zhang, J. Chen, Y. Zhang, S. An, Yb^{3+}/ Tm^{3+} and Yb^{3+}/Ho^{3+} doped $NaY_9(SiO_4)_6O_2$ phosphors: Upconversion luminescence processes, temperature-dependent emission spectra and optical temperature-sensing properties, J. Alloys Compd. **860** (2021). https://doi.org/10.1016/j.jallcom.2020.158473.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

2024 IEEE 9th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 5-7, 2024

Visual Sentiment Recognition via Popular Deep Models on the Memotion Dataset

Bhavana Verma
Biometric Research Laboratory, Department of Information Technology, Delhi Technological University
Bawana Road, Delhi-110042, India
bhavana.verma2905@gmail.com

Priyanka Meel
Assistant Professor, Department of Information Technology, Delhi Technological University
Bawana Road, Delhi-110042, India
priyankameel86@gmail.com

Dinesh Kumar Vishwakarma
Biometric Research Laboratory, Department of Information Technology, Delhi Technological University
Bawana Road, Delhi-110042, India
dvishwakarma@gmail.com

2024 IEEE 9th International Conference for Convergence in Technology (I2CT) | 979-8-3503-9447-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/I2CT61223.2024.10543569

***Abstract*—Sentiment analysis has historically concentrated on text-based data to investigate people's views towards objects. However, given how frequently visual content is shared on social media, a strong case exists to expand sentiment analysis to include image context. This research addresses the gap through a novel approach by comparing several deep learning methods—such as CNNs and transfer learning—on the Memotion dataset. Our approach is unique in that we carefully analyze performance indicators, paying particular attention to feelings associated with the offensive class and emotions in the images. Our study explores the subtleties of sentiment analysis in visual content by assessing accuracy, precision, recall, and F1-score.By evaluating accuracy, precision, recall, and F1-score, our study delves into the nuanced aspects of sentiment analysis in visual content. Notably, we contribute valuable insights by comparing ResNet, DenseNet, EfficientNet, and Swin Transformer models, revealing that DenseNet emerges as the superior choice, showcasing advancements in image sentiment analysis techniques. This comparative analysis not only highlights the effectiveness of different models but also underscores the innovation in our methodology, providing a significant contribution to the evolving landscape of sentiment analysis in the visual domain.**

Keywords—Sentiment analysis, Emotion recognition, ResNet,DenseNet,Swin Transformer,EfficientNet.

I. INTRODUCTION

Sentiments are feelings such as love, grief, happiness, and anger that impact others through emotion, and sentiment analysis categorizes them into positive, negative, or neutral classes[1]. Policymakers, for example, can utilize feelings to monitor public attitudes toward policies. The hotel staff uses client feedback to make improvements to the facilities. That is why emotional awareness is essential for understanding.

Traditionally, we may anticipate sentiments based on unimodality[2], the bulk of which is text in which the relationship between words and phrases is analyzed. Still, text alone is insufficient to extract people's feelings or sentiments. For example, in analyzing text sentiment, the word "amazing" is normally favorable, but when combined with "well," it signals doubt or sarcasm, indicating negative attitudes. To overcome the limitation of traditional text-based sentiment analysis, multimodal sentiment analysis[3] is offered, comprising a combination of two or three modalities with a massive quantity of social network data available in video, image, or voice.

People increasingly share their views and emotive memeson various platforms, including Instagram, WhatsApp, YouTube, and others, providing excellent data for sentiment analysis. Memes are new messages created by a caption on an image or text inserted into an image[4]. It can be an amusing notion or behavior that passes from person to person, but it can also be adverse in the context of sharing someone's private images without consent, misleading people by providing incorrect political informationor bullying them by forwarding it[5]. Consider a meme that includes the caption "she's fine" but in the image a woman is drunk and two men are helping her, which simply shows the sarcasm, thus meme analysis is important for growing research and we need sentiment analysis to precisely identify memes.Because it can be both a visual and textual representation, a meme is particularly multimodal. It can also be applied in a multimodal setting, where the memewith or without textis combined with textual remarks and exchanges to offer emphasis, more context, or an emotive reaction[6].The relationship between the two items may be used to determine the meaning of the word when an image of an apple is included to go along with the text. Within the framework of this essay, the meme is comparable to the apple picture, with the term "apple" standing in for social media communications. The meme offers more information on the text's meaning or response to it[7].At first, it was believed that the emotive level was subjective and that multimedia picture retrieval was based on real or objective data.As effective multimedia discovery developed further, it became clear that effective content could also be objective rather than subjective if the effect was a predictable outcome of the designer's intended actions or if the multimedia content elicited a widely shared affective response in society[8].

Fig 1: Visualization of Text in image, commonly known as Meme

Textual content analysis includes word, phrase, remarks, and discussion analysis. Images were categorized by use of pictures. In addition, if textual messages were embedded in the memes, they were identified and evaluated[6].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:18:46 UTC from IEEE Xplore. Restrictions apply.

Various text and image-based deep learning methods are used for sentiment analysis. For detecting the linguistic patterns in text various supervised and unsupervised learning techniquesare used for the successful analysis of sentiments [9].We can use Early Fusion, Late Fusion,and Attention-Based Cross-Modal[10] models like CNN[11], RNN, and Transformers for images and LSTM[12], Bi-LSTM, and BERT for text sentiment analysis.

The main contribution of our work is that, since most work on this dataset focuses on CNN, BERT, and ResNet, we are offering a novel way for the Memotion dataset to compare multiple deep-learning approaches, such as ResNet18, ResNet50, DenseNet121, EfficientNet, and Swin Transformer. This allows us to achieve better results on various evaluation metrics using different transform learning techniques.Accuracy, Precision, Recall, and F1 score are just a few examples of the metrics that can be used to evaluate thedataset.

The remaining manuscript is structured as follows. The majority of the sentiment analysis study is highlighted in Section II. The procedure is described in Section III. The experimental details are covered in Section IV, and the study is wrapped up in Section V.

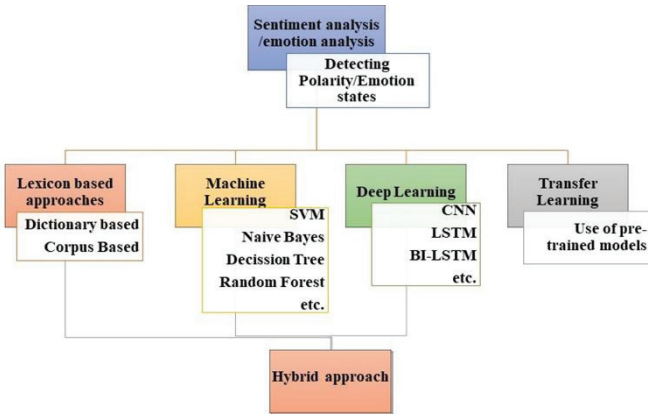


Fig 2: Techniques of Sentiment Analysis[13]

II. Related Work

This section represents the literature survey of deep learning techniques used for classifying meme emotions. Memes are spreading humor to target individuals or groups of people in a sarcastic or polite way, it is very difficult to classify memes.

Aggarwal et al. [14] proposed a model that can classify more than one modal by using two different approaches i.e., sequential and multimodal where the first approach classifies text and image sentiments and the second approach integrates the features of different modalities.Sherratt et al. [15] proposed an MC-CNN model for classifying memes in live data and wild data. To provide accuracy in data they used a Multi-channel system i.e. Text, image, and histogram which ranks the non-meme formats with textual attributes. Agarwal and Meel[16] proposed a deep learning framework using stacked BI-LSTM which incorporates embeddings from fine-tuned BERT model and provides highest accuracy of 0.99 on the Articles dataset and Kaggle dataset.

Barveen et al.[17] proposed a method of Extended LSTM-RNN which provides better results as compared to CNN and analyses features using HOG feature extraction which divides the image into smaller segments that classify the image as a good meme or not with an accuracy of 0.98%. Tang et al.[18]proposed a multimodal bi-directional dynamic-enhanced block that will reduce the redundancy and provide fine-grained auxiliary context modalities on the MOSI dataset. Intra-modality redundancy between the audio and video block is reduced by using this model which provides fine-grained sentiment contexts. Xue et al.[19] proposed a MAMN model whose performance is compared on the basis of 19 methods and provides noise reduction on multilevel granular data in an efficient way. Before reducing the noise, it first, extracts the granular features and then applies fusion methods for document and aspect-based sentiment analysis.

Zhu et al.[20] proposed a cross-modal ITIN network to capture the finer relationship between text and image on multimodal features for classifying the sentiments on the MVSA dataset. Zeng et al. [21] proposed a Gated Fusion Semantic Network to provide the inter and intra relationship between text and visual modalities. A cross-modal mechanism improves the result by 0.10% to 3.66% by employing cross-attention between the text embedding and images for the sentiment prediction on the Yelp and Flickr30k datasets. Meel and Kumar Vishwakarma[22] proposed a frameworkthat uses multi-modal fusion architecture for both textual and visual data and achieves 97.19% accuracy on Weibo[23], and the MediaEval16 dataset.

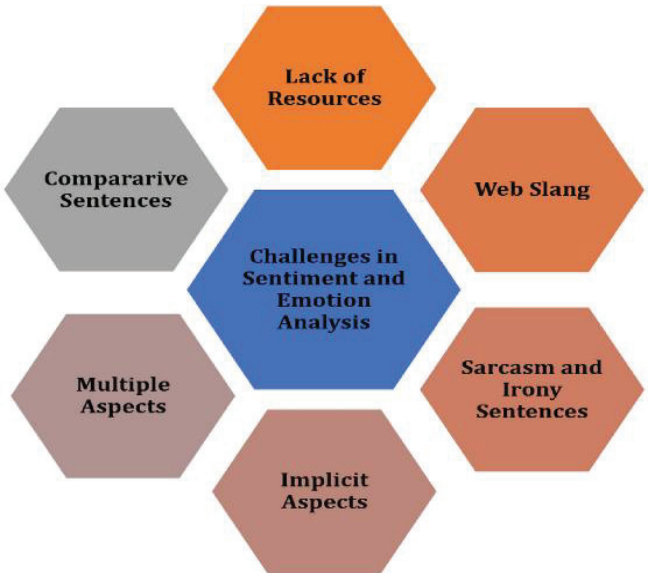


Fig 3: Challenges in sentiment analysis[13]

Yao et al.[24] proposed a sarcasm detection model M3N2 for mimicking the brain's perception using guided attention and attention pooling methods for capturing the modal information on Twitter images, Text embedded on images, and caption of tweets from different levels like singlewise, pairwise, triplewise, and tetradwise. Stappen et al.[25] proposed a multimodality solution that uses the MAHN modelto categorize the emotion from MuSe 2020 dataset which contains features in the form of Audio, Video, and text. Ye et al.[26] proposed a SMP framework for addressing the sentiment signals, it captures theinformation from text and images using cross-modal contrastive modules on VSO, MVSA, Twitter15, and Twitter17 datasets and provides fine-grained sentiment analysis. Zeng et al.[27] proposed a TATE network for handling the missing partially

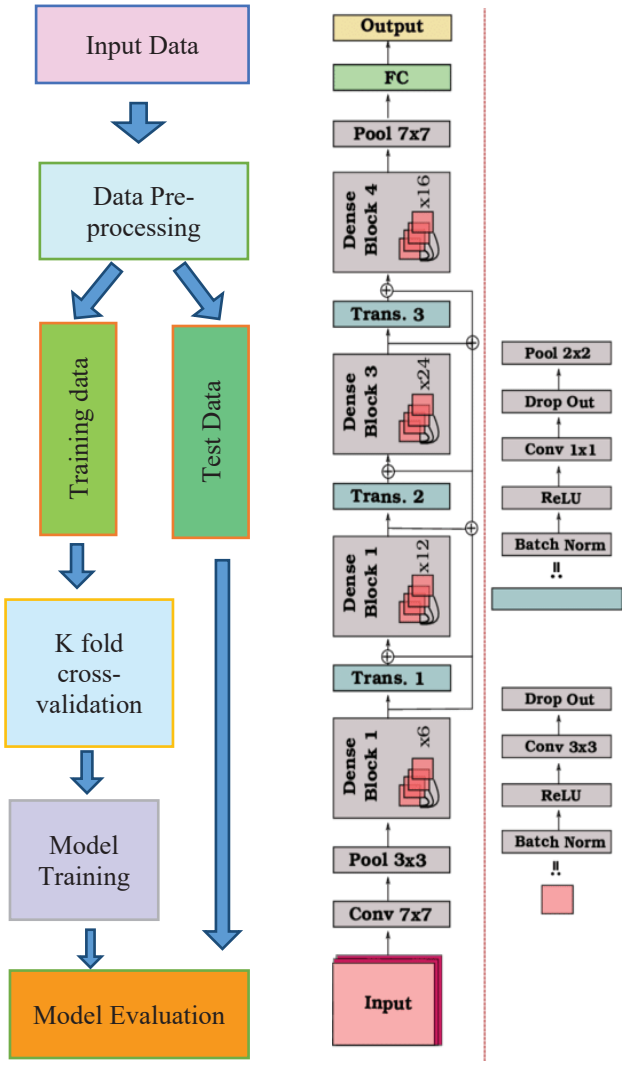
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:18:46 UTC from IEEE Xplore. Restrictions apply.

available multiple modalities instead of considering only the single modality. To learn the missing modality features, aligned vectors are fed into the transformer and then transformer encoder is used on CMU-MOSI and IEMOCAP datasets for classifying the sentiments.

After thoroughly survey the papers we conclude that there is a need of comparative study about deep learning methods on Memotion dataset.

III. Methodology

In this Section, we provide details about deep learning methods and explain the working of our model and their key components.



(a) (b)

Fig 4: (a)Procedure flow diagram for the proposed approach (b) Basic architecture of DenseNet-121

A. Overview

Our approach begins with the initial stage of processing the data, which is then followed by a phase of data pre-processing. Data pre-processing is essential for cleansing the image data, assuring its integrity, and preserving its quality. Afterward, the dataset is partitioned into several subsets for training and testing. The training data is further adjusted using k-fold cross-validation. Once the data is in optimal condition, several deep learning networks such as CNNs, RNNs, and Transfer Learning methods are employed to evaluate metrics such as accuracy and precision. The outcomes can be visually represented using a variety of charts and graphs to evaluate the effectiveness of the model. This helps to figure out which method or model works best for the task. Figure 4(a) shows the flow diagram for the proposed approach.

B. Image Models:

- ResNet: A residual network is used to solve the gradient problem that arises due to an increase in the number of layers in CNN. The skip connections technique is used in which the activation connection of a layer to another layer is done by skipping some layers between them, which is known as a residual block. Residual networks are inspired by VGG-19, which uses a 34-layer plain network and a shortcut connection to transform it into a residual network that improves the performance of deeper-layer networks[28].

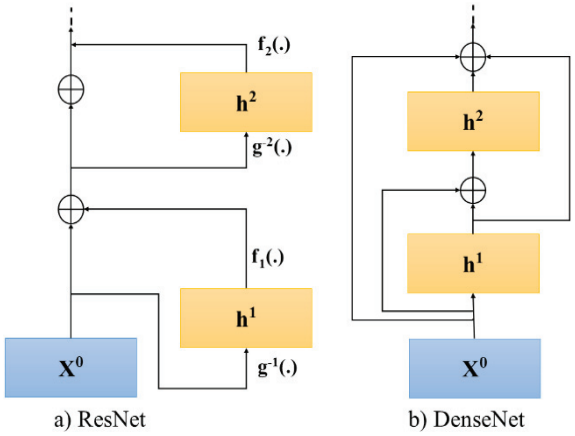


Fig5: ResNet and DenseNet

- DenseNet: Densely connected-convolutional network directly connects any layer to all subsequent layers unlike ResNet in which skip-connection is used to pass information from one layer to another layer[29].
- EfficientNet: DenseNet balances the depth of the network, for balancing the depth, width and resolution of the network we uses the EfficientNet[30].
- SwinTransformer: It is used for image classification and object detectionand provides better performance thanother vision networks like CNNs and ViTs. Its key features are first it is designed in hierarchical structure stages where each stage process image at different level. Secondly, it uses shifted windows to process image patches, then for creating relationships between these image patches multi-head-self-attention is used[31].

C. Abbreviations

DenseNet without transform (DN121), Resnet50 with transform (RN50 T), Resnet50 without transform (RN50), Resnet18 without transform (RN18), Resnet18 with transform (RN18 T), Efficienet without transform (EF), SwinTransformer without transform (ST).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:18:46 UTC from IEEE Xplore. Restrictions apply.

IV. Experimental Details

A. Dataset

This online memedataset contains approx. 10K memes having labels such as sentiments based on (positive, negative, and neutral) and type of emotions such as humor, sarcasm, offensive and motivational.

Memotion dataset[32]are classified into different emotions such as humor, sarcasm, offensive and motivational and based on this an overall sentiment is labeled as positive, negative and neutral. For our experiment we are trying to classify memes into 4 levels of offensive class which are non-offensive, slight, very offensive and hateful offensive. Among these 10k samples of memes 80% samples uses as train data and 20% as test data. The percentage of data classification in offensive classes are mentioned in Table 1.

TABLE I. Offensive Class Percent Classification

Offensive Classes	Percentage in Data
Very offensive	20.97%
No offensive	38.80%
Hateful offensive	3.16%
Slight Offensive	37.07%

B. Data Pre-processing

In the data preprocessing phase, we implement a series of essential steps to enhance the quality and usability of our image dataset. Initially, we undertake a cleaning process, systematically removing unlabeled images to ensure the integrity of our dataset. Subsequently, we rescale the images to a standardized 128x128 resolution, facilitating consistent and efficient processing. To augment the dataset and diversify the training samples, we employ a range of transformation techniques. Some of the common operations that can be performed include horizontal and vertical flips, shift-scale rotations, RGB shifts, resizing, and random adjustments to brightness and contrast. These transformations enhance the learning of our dataset by using deep learning models.

C. Hyperparameters Tuning

To achieve better model performance hyperparameter tuningmethods are used. For the integration of 5-fold cross-validation with more than 100 epochs used to explore the model learning capability with a learning rate of 0.01 to achieve the highest accuracy. The implementation of all models is conducted using the PyTorch framework, to effectively utilize the computational power of Nvidia GPU. The detailed description of the models along with their parameters and accuracy are defined in Table 2.

D. Evaluation Metrics

In this section,a set of key assessment measures for the comprehensive comparative analysis of a given dataset is mentioned.Precision [1], Recall [2],F1-score [3], and Accuracy [4] are the assessment metrics used for the comparative analysis of this image dataset.Precision[33], calculated as the ratio of true positives to the sum of true positives and false positives highlights the model's accuracy

TABLE II. Model Description

Model	Parameterization	Acc. Scores
ResNet18	The model is pretrained on the IMAGENET1K_V1 dataset with the training epochs (100), batch size (500), learning rate (0.01), and the use of 5-fold-cross validation.	**56%**
ResNet50	We are getting better results using SGD optimizer with 5-fold-cross-validation and having epochs 100 with learning rate is 0.01.	**49%**
DenseNet121	The model process on image size of 128x128 having batch size 500 with learning rate 0.01 for SGD optimizer.	**95%**
EfficientNet	With batch size of 500, a learning rate of 0.01. The model was trained for 100 epochs using fivefold cross-entropy loss function. The model accuracy was discovered after third fold.	**49%**
Swin Transformer	For this model we use a batch size of 250 with 100 epochs using five folds and provides better accuracy after third fold for SGD optimizer.	**43%**

in identifying positive instances. Similarly, Recall[33], is determined by the ratio of true positives to the sum of true positives and false negatives which showcases the model's ability to capture all relevant positive instances.Accuracy[34], defined as the ratio of correctly predicted instances (both true positives and true negatives) to the total instances, provides an overall assessment of model correctness. The F1-score is a harmonic mean of Precision and Recall.By employing these standard metrics, we ensure a comprehensive and rigorous evaluation of the model's performance on the image dataset.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (4)$$

V. Results& Analysis

The result section provides a detailed comparative analysis of various deep learning models applied on the Memotion dataset. Table 3 offers an in-depth examination of the performance metrics, including accuracy, precision, recall, and F1 score, highlighting the subtle distinctions among ResNet, ResNet50, EfficientNet, Densenet121, Swin Transformer, and other models. Insightful observations emerge from our study, prominently illustrated in Figure6, where the comparative study graph showcases DenseNet's exceptional performance, achieving an accuracy of 95%, surpassing other vision transformer models. This superiority of DenseNet comes from its unique architecture which provides densely connected blocks that effectively addresses issues like the vanishing gradient problem and promote

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:18:46 UTC from IEEE Xplore. Restrictions apply.

efficient features reuse.DenseNet121 consists of 120 convolutional layers and 4 average pooling layers as shown in Figure 4(b). Within each dense block, the weights are distributed across multiple inputs, enabling deeper layers to utilize the extracted features. In the second and third dense blocks, the output of the transition layers is given the lowest weight assignment. Even though the final layers use the weights of the whole dense block, more high-level features

may be created further into the model. This was seen in the tests, where there was more focus on the final feature maps. Due to its reduced parameter requirement and feature reuse capability, DenseNet yields more compact models in comparison to CNN or ResNet.

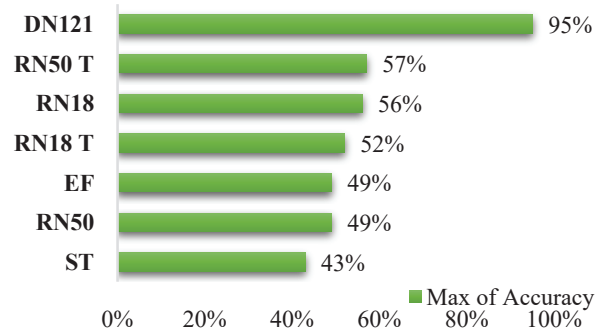


Fig 6: Accuracy graph for comparative models

Figure7 extends the comparison to different vision models on the IMAGENET1K_V1 pretrained dataset, visually representing the performance differences among Resnet18, Resnet 18 with Transform, RN50, Efficientnet, Densenet121, and Swin Transformer. Particularly, DenseNet stands out across multiple metrics, establishing its advantages in image sentiment analysis. Figure8 illustrates the positive correlation between the accuracy of the model and the amount of data provided to the model. This improvement in accuracy is achieved through the use of the five-fold cross-validation technique and the application of various transformers during the data pre-processing phase. These findings emphasize the durability and effectiveness of DenseNet, confirming its superiority over alternative models for sentiment analysis tasks using visual data.

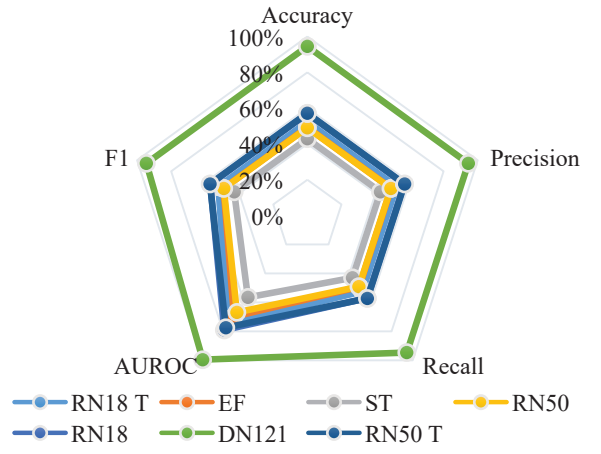


Fig7: Comparison between different models according to the assessment metrics

TABLE III. EXPERIMENTAL RESULTS

Model Name	Accuracy	Precision	Recall	F1
DN121	94.8%	94.6%	94.2%	94.0%
RN50 T	57.0%	56.1%	55.6%	55.8%
RN18	56.0%	56.2%	56.1%	56.3%
RN18 T	52.0%	51.6%	51.2%	51.3%
EF	48.6%	48.0%	47.0%	47.5%
RN50	48.5%	49.0%	47.2%	48.0%
ST	42.8%	42.0%	43.0%	42.3%

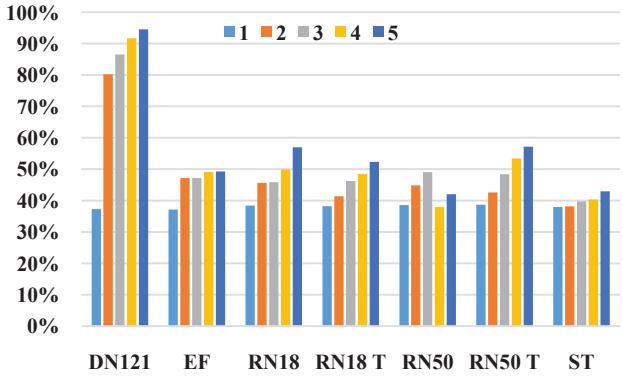


Fig 8: Accuracy enhancement as compared to No. of folds

VI. CONCLUSION&FUTURE WORK

Thecomparative analysis of various deep learning methods proves instrumental in the informed selection of techniques tailored to specific dataset limitations, systematically unveiling the strengths and weaknesses inherent in each approach. The insights gained from this comparison not only aid in the current methodological choices but also guide future research endeavours. By identifying effective adjustments or novel combinations of techniques, the study contributes to the evolving landscape of sentiment analysis in the visual domain. Particularly noteworthy is the experimental finding that DenseNet outperforms other deep learning techniques on the Memotion dataset, achieving an impressive 95% accuracy compared to Swin Transformer and EfficientNet. This result highlights the efficacy of DenseNet for image sentiment analysis.Looking ahead, our methodology offers a versatile framework that can be extended to text models, image models, and their fusion, promoting the exploration of multimodal approaches for a more comprehensive understanding of sentiments and emotions.

Future research should utilize Densenet121's exceptional performance on the Memotion dataset by enhancing its capabilities through targeted transfer learning and fine-tuning. Looking into how interpretable Densenet121 is in meme analysis might help us understand more about the most important visual features, which would improve the overall interpretability of models in meme-related research. Furthermore, a study might be devised to examine the dissemination of memes throughout various communities by assessing the level of popularity or aversion towards a specific meme inside a particular community, such as

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:18:46 UTC from IEEE Xplore. Restrictions apply.

Facebook, Instagram, WhatsApp, and others. This investigation expands to reveal political tendencies within memes, providing insights into the attitudes conveyed by different groups and contributing to a sophisticated comprehension of meme's influence on contemporary political communication.

REFERENCES

[1] G. T. Kannan, M. Gunasekar, N. A. Ponnazhagan, K. R. Prasanna Kumar, S. Sudhan, and K. Logeswaran, "Aspect Based Sentiment Aware Word Embedding for Cross Domain Sentiment Analysis," in 2023 International Conference on Computer Communication and Informatics (ICCCI), Jan. 2023, pp. 1–5. doi: 10.1109/ICCCI56745.2023.10128251.

[2] S. Peng et al., "A survey on deep learning for textual emotion analysis in social networks," Digit. Commun. Netw., vol. 8, no. 5, pp. 745–762, Oct. 2022, doi: 10.1016/j.dcan.2021.10.003.

[3] B. Yang, L. Wu, J. Zhu, B. Shao, X. Lin, and T.-Y. Liu, "Multimodal Sentiment Analysis With Two-Phase Multi-Task Learning," IEEEACM Trans. Audio Speech Lang. Process., vol. 30, pp. 2015–2024, 2022, doi: 10.1109/TASLP.2022.3178204.

[4] N. Saini, M. Singhal, M. Tanwar, and P. Meel, "Multimodal, Semi-supervised and Unsupervised web content credibility analysis Frameworks," in 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), May 2020, pp. 948–955. doi: 10.1109/ICICCS48265.2020.9121005.

[5] D. Morina and M. S. Bernstein, "A Web-Scale Analysis of the Community Origins of Image Memes," Proc. ACM Hum.-Comput. Interact., vol. 6, no. CSCW1, p. 74:1-74:25, Apr. 2022, doi: 10.1145/3512921.

[6] J. H. French, "Image-based memes as sentiment predictors," in 2017 International Conference on Information Society (i-Society), Jul. 2017, pp. 80–85. doi: 10.23919/i-Society.2017.8354676.

[7] B. Adams, D. Phung, and S. Venkatesh, Funniest thing i've seen since [href="http://flic.kr/p/KGEGB"]: Shifting perspectives from multimedia artefacts to utterances. 2012, p. 60. doi: 10.1145/2390876.2390894.

[8] A. Hanjalic, "Adaptive extraction of highlights from a sport video based on excitement modeling," IEEE Trans. Multimed., vol. 7, no. 6, pp. 1114–1122, Dec. 2005, doi: 10.1109/TMM.2005.858397.

[9] D. K. Vishwakarma, P. Meel, A. Yadav, and K. Singh, "A framework of fake news detection on web platform using ConvNet," Soc. Netw. Anal. Min., vol. 13, no. 1, p. 24, Jan. 2023, doi: 10.1007/s13278-023-01026-7.

[10] X. Zou, S. Wu, N. Zhang, and E. M. Bakker, "Multi-label modality enhanced attention based self-supervised deep cross-modal hashing," Knowl.-Based Syst., vol. 239, p. 107927, Mar. 2022, doi: 10.1016/j.knosys.2021.107927.

[11] A. Verma, V. Ranga, and D. K. Vishwakarma, "A novel approach for forecasting PM2.5 pollution in Delhi using CATALYST," Environ. Monit. Assess., vol. 195, no. 12, p. 1457, Nov. 2023, doi: 10.1007/s10661-023-12020-z.

[12] A. Verma, V. Ranga, and D. K. Vishwakarma, "Forecasting of Satellite Based Carbon-Monoxide Time-Series Data Using a Deep Learning Approach," in 2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT), Feb. 2023, pp. 1–4. doi: 10.1109/ICITIIT57246.2023.10068609.

[13] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," Soc. Netw. Anal. Min., vol. 11, no. 1, p. 81, Aug. 2021, doi: 10.1007/s13278-021-00776-6.

[14] A. Aggarwal et al., "Two-Way Feature Extraction Using Sequential and Multimodal Approach for Hateful Meme Classification," Complexity, vol. 2021, p. e5510253, Apr. 2021, doi: 10.1155/2021/5510253.

[15] V. Sherratt, K. Pimbblet, and N. Dethlefs, "Multi-channel Convolutional Neural Network for Precise Meme Classification," in Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, in ICMR '23. New York, NY, USA: Association for Computing Machinery, Jun. 2023, pp. 190–198. doi: 10.1145/3591106.3592275.

[16] A. Agarwal and P. Meel, "Stacked Bi-LSTM with Attention and Contextual BERT Embeddings for Fake News Analysis," in 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Mar. 2021, pp. 233–237. doi: 10.1109/ICACCS51430.2021.9441715.

[17] A. Barveen, S. Geetha, and M. K. M. Faizal, "Meme Expressive Classification in Multimodal State with Feature Extraction in Deep Learning," in 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Apr. 2023, pp. 1–10. doi: 10.1109/ICEEICT56924.2023.10157066.

[18] J. Tang et al., "BAFN: Bi-Direction Attention Based Fusion Network for Multimodal Sentiment Analysis," IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 4, pp. 1966–1978, Apr. 2023, doi: 10.1109/TCSVT.2022.3218018.

[19] X. Xue, C. Zhang, Z. Niu, and X. Wu, "Multi-Level Attention Map Network for Multimodal Sentiment Analysis," IEEE Trans. Knowl. Data Eng., vol. 35, no. 5, pp. 5105–5118, May 2023, doi: 10.1109/TKDE.2022.3155290.

[20] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal Sentiment Analysis With Image-Text Interaction Network," IEEE Trans. Multimed., vol. 25, pp. 3375–3385, 2023, doi: 10.1109/TMM.2022.3160060.

[21] J. Zeng, J. Zhou, and C. Huang, "Exploring Semantic Relations for Social Media Sentiment Analysis," IEEEACM Trans. Audio Speech Lang. Process., vol. 31, pp. 2382–2394, 2023, doi: 10.1109/TASLP.2023.3285238.

[22] P. Meel and D. Kumar Vishwakarma, "Multi-modal fusion using Fine-tuned Self-attention and transfer learning for veracity analysis of web information," Expert Syst. Appl., vol. 229, p. 120537, Nov. 2023, doi: 10.1016/j.eswa.2023.120537.

[23] "Papers with Code - DeepHawkes: Bridging the gap between prediction and understanding of information cascades." Accessed: Nov. 06, 2023. [Online]. Available: https://paperswithcode.com/paper/deephawkes-bridging-the-gap-between

[24] F. Yao, X. Sun, H. Yu, W. Zhang, W. Liang, and K. Fu, "Mimicking the Brain's Cognition of Sarcasm From Multidisciplines for Twitter Sarcasm Detection," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 1, pp. 228–242, Jan. 2023, doi: 10.1109/TNNLS.2021.3093416.

[25] L. Stappen, A. Baird, L. Schumann, and B. Schuller, "The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements," IEEE Trans. Affect. Comput., vol. 14, no. 2, pp. 1334–1350, Apr. 2023, doi: 10.1109/TAFFC.2021.3097002.

[26] J. Ye et al., "Sentiment-aware multimodal pre-training for multimodal sentiment analysis," Knowl.-Based Syst., vol. 258, p. 110021, Dec. 2022, doi: 10.1016/j.knosys.2022.110021.

[27] [J. Zeng, T. Liu, and J. Zhou, "Tag-assisted Multimodal Sentiment Analysis under Uncertain Missing Modalities," in Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid Spain: ACM, Jul. 2022, pp. 1545–1554. doi: 10.1145/3477495.3532064.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[29] [G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[30] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." arXiv, Sep. 11, 2020. Accessed: Aug. 15, 2023. [Online]. Available: http://arxiv.org/abs/1905.11946

[31] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." arXiv, Aug. 17, 2021. doi: 10.48550/arXiv.2103.14030.

[32] C. Sharma et al., "SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor!," in Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 759–773. doi: 10.18653/v1/2020.semeval-1.99.

[33] "Classification: Precision and Recall | Machine Learning," Google for Developers. Accessed: Nov. 21, 2023. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:18:46 UTC from IEEE Xplore. Restrictions apply.

[34] "Evaluation Metrics for Machine Learning - Accuracy, Precision, Recall, and F1 Defined," Pathmind. Accessed: Nov. 21, 2023. [Online]. Available: http://wiki.pathmind.com/accuracy-precision-recall-f1

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 18,2024 at 09:18:46 UTC from IEEE Xplore. Restrictions apply.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19582-1

Weighted voting ensemble of hybrid CNN-LSTM Models for vision-based human activity recognition

Sajal Aggarwal[1] · **Geetanjali Bhola**[1] · **Dinesh Kumar Vishwakarma**[1]

Received: 25 April 2023 / Revised: 22 May 2024 / Accepted: 30 May 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
This research work aims to propose an ensemble model of different pre-trained CNN networks combined with LSTM to detect a set of routine human activities practiced by the elderly from the publicly available STAIR Actions Dataset. The prime objective of this work is to visualize the effect of using different networks for implementing activity detection for a set of five activities, compare them based on standard performance parameters and combine the best- performing models into a better-performing ensemble unit. Applying the transfer learning technique helped achieve shorter training times and promising classification results for the dataset. The hybrid CNN-LSTM model uses CNN to extract spatial features from the video frames, followed by classification using sequential learning by the LSTM network. For the set of five routine activities practised by the elderly taken up from the STAIR Actions dataset, we obtained an overall test set accuracy of 82.27%, 82.73%, 73.64%, 72.73%, 64.55%, 71.82%, 83.64%, 88.18%, 86.82% and 86.82% respectively on using VGG16, VGG19, ResNet50, ResNet101, ResNet152, Inception V3, Xception, DenseNet121, DenseNet169 and DenseNet201 models, pre-trained on the ImageNet dataset for feature extraction from the video frames. Apart from test-set accuracy, we evaluated the models using other performance metrics, including AUROC, AUPRC, sensitivity, specificity and precision. Using the best-performing model from each category of CNN networks employed, we developed a weighted voting ensemble that outperformed the best-performing individual model in terms of test set accuracy. The highest test set accuracy achieved by optimising the weights assigned to each model in the proposed ensemble was 93.18%.

Keywords Vision-based human activity recognition · LSTM · CNN · Ensemble Model · Video Classification · Deep Learning · Neural Networks · Weighted Voting

✉ Dinesh Kumar Vishwakarma
dvishwakarma@gmail.com

Sajal Aggarwal
authorsajal@gmail.com

Geetanjali Bhola
geetanjali.getz@gmail.com

[1] Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Bawana Road, Delhi 110042, India

Published online: 08 June 2024

1 Introduction

Modern-day applications of the advancements in 2D and 3D video technologies such as intelligent surveillance [1, 2] and security systems or smart home [3, 4] and healthcare systems [5, 6] have one aspect in common, i.e., Human Activity Recognition (HAR). HAR is one of the most crucial components in enabling systematic and personalized healthcare and life care systems. The ability to successfully monitor and track human activities correctly can make providing the elderly with premium quality healthcare services much easier and also help save a considerable amount of resources that would otherwise be required for healthcare expenditure. Such services can help prolong the life of the elderly by taking good care of them.

With the number of older people living by themselves increasing manifold, the use of human activity recognition to develop assistive robots could keep track of their daily activities [7, 8] and help them perform necessary activities such as taking medication at the right time, which by old age they themselves tend to forget. The ability to understand human activities and respond appropriately can completely transform robotic science to an entirely different level; as such, robots would no longer require instructions to perform tasks. They would rather be able to perform a follow-up task in response to an activity they detect, thus integrating them with human interactions.

With the emergence of the COVID-19 pandemic, people infected with the virus had to be isolated or home-quarantined to avoid further spread. On account of the more significant health risk faced by them, the elderly were subjected to complete home confinement to avert any chances of infection. This confinement, coupled with an acute shortage of healthcare staff raised the need to track their everyday activities [9] using vision-based methods to monitor their health. Keeping this scenario in mind, the activities chosen for this work are routine activities practised by the elderly in indoor environments.

Detection of human activities accurately can prove to be quite a challenging task for several reasons, including the presence of several other unnecessary objects or activities taking place in a particular frame captured to detect a particular activity and ambiguity in detection. Most human activities, such as drinking water or eating a meal, can be performed in a wide variety of body poses or orientations. While detecting activities, we come across inter-activity (this means that two or more activities being detected involve a somewhat similar body movement, thus creating confusion in detection) and intra-activity (these account for the differences that arise due to the same activity being performed differently by different subjects used for testing and training) variations [10, 11]. This makes video-based activity recognition a somewhat cumbersome task.

HAR has been carried out primarily using two broad approaches - sensor-based [12] and vision-based [13–16]. Wearable sensor-based HAR involves multiple sensors that are continually affixed to a person's body. The severable cable connections may prove to be uncomfortable for the user, along with other disadvantages such as high costs, energy consumption, and device setup. The dependability on sensors gets obliterated if we directly use the camera feed to detect activities using vision-based HAR. Additionally, adding activities to the vision-based model is easier without additional hardware requirements. Though, compared to the sensor-based methods, an elaborate setup is not required for vision-based detection before implementation and sensor cost gets eliminated, vision-based activity recognition has its own set of challenges, as discussed in detail in [14, 17].

According to the feature extraction process, vision-based human activity recognition methods can be classified into handcrafted and learning-based processes. Due to the

increased computational power required by deep learning-based approaches to process sequential image frames representing human actions, the field of activity recognition was initially dominated by handcrafted action representations. These methods involve three major steps – foreground detection, feature selection and extraction and classification of actions represented by extracted features using a generic trained classifier [18]. The feature learning processes can be classified into traditional and deep learning methods. Deep-learning methods are now being widely explored for activity recognition on account of their promising results and generalization ability for different types of data. Contrary to handcrafted feature extraction, these methods involve an automated feature extraction [13]. The primary challenges posed by these methods comprise the requirement of huge amounts of data to act as a training set and the increased computational complexity involved. The steps involved in deep learning methods for vision-based human activity recognition have been shown in Fig. 1.

1.1 Aim of the study and major contributions

This study aims to experiment with multiple cutting-edge CNNs to develop hybrid CNN-LSTM models for vision-based HAR and combine the front-running networks via weighted voting to generate an improved ensemble model. The major contributions of this paper are summarized as follows:

(a) Implementing a hybrid model that capitalizes on the strengths of state-of-the-art CNN and LSTM networks to efficiently encapsulate both the spatial and long-range temporal dependencies for vision-based HAR.
(b) Extensive comparison of ten hybrid CNN-LSTM models using various performance criteria such as precision, specificity, sensitivity, accuracy, AUROC, and AUPRC.
(c) A conclusive model that outperforms previous baselines in terms of test-set accuracy. This is achieved by assembling the top-performing hybrid models into a composite ensemble model. The final model predictions are determined through weighted voting of individual judgements, with the optimal weights established through an exhaustive grid search technique.

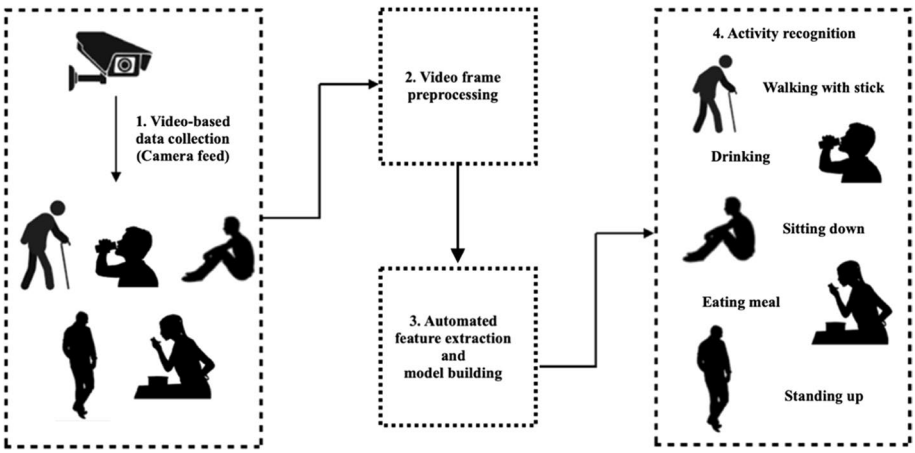


Fig. 1 Fundamental steps involved in deep learning approaches to vision-based human activity recognition

(d) Ablation study to further strengthen the experimental findings and confirm the importance of LSTM in the hybrid model.

1.2 Article organization

In the following sections, we discuss our aim and approach in more detail and provide further information. The rest of this paper is organized as follows: Section 2 highlights the related work carried out in the field of human activity recognition. Section 3 explains the proposed approach. Section 4 presents the experimental setup to conduct a detailed analysis and review of all individual models trained and the ensemble model proposed. Section 4.5 presents the results achieved using the individual hybrid models and the ensemble model. Section 5.6 delineates the merits and demerits of the proposed architecture. Finally, Section 7 concludes the paper and provides suggestions for future works.

2 Related work

With the tremendous amount of focus it has gained, HAR has acquired a very vast literature. Primarily, activity recognition has been performed using either sensor or visual data.

2.1 Sensor-based Human activity recognition

This is a crucial research field that focuses on wearable or smartphone-embedded sensors such as accelerometers, gyroscopes, GPS, and temperature sensors to identify and understand human activities [19–23]. Sri Harsha et al. [24] compared the performance of machine learning algorithms for smartphone-based HAR. Raj and Kos [25] proposed a multi-layered CNN architecture for activity recognition. Recurrent neural networks, including LSTMs, have been widely used for sensor-based HAR [26–30]. Kaya and Topuz [31] introduced a novel 1-D CNN that utilizes smartphone sensor data to identify human activities. Park et al. [32] proposed a novel grouped temporal shift network (GTSNet) that is designed to be adaptable by establishing the computing cost to align with budgetary limits on different wearable devices. Matey-Sanz et al. [33] curated a novel dataset to offer smartwatch and smartphone-based inertial data for sensor-based HAR. The data samples were gathered from a diverse range of individuals in terms of age and gender. Zhang and Yu et al. [34] proposed integrating a 1-D CNN with a BiLSTM model and an attention mechanism to enhance multi-sensor HAR. Ullah and Pirahandeh et al. [35] put forth a novel deep-learning architecture for sensor-based HAR. The authors employed sparse learning to retrain just the weak nodes and keep stronger nodes fixed in a fully linked layer before the final classification. By integrating CNN, BiGRU, and BiLSTM networks into a single architecture, Lalwani and Ramasamy [36] aimed at capturing both short-term and long-term patterns in sequential HAR data. Sansano et al. [37] experimented with several state-of-the-art deep-learning architectures such as CNN, LSTM, BiLSTM, GRU, and DBN for HAR using different types of sensors. Mallik et al. [38] presented a dual-phase approach to categorize activities into static and dynamic classes. While the first phase was based on SVM, the second used 2D and 1D CNN to identify static and dynamic activities. For sensor-based HAR, Serpush et al. [39] employed CNN and fuzzy SVM as the feature extractor and classifier, respectively. Noori et al. [40] explored different methods for integrating multiple

sources of sensor data for HAR. The authors compared the performance of data, feature, and decision-level fusion methods across three benchmark datasets. After extracting relevant features from sensor data using a fuzzy logic-based genetic algorithm, Jameer and Syed [41] used a CNN-LSTM framework to classify human activities. To combat the effects of defective sensors and missing data, Bagci Das and Birant [42] devised a flexible multi-instance learning framework for sensor-based HAR. Kumar and Murugan [43] suggested using a deep belief network for classification of sensor data while performing hyperparamter tuning using the cuckoo search optimization algorithm. Choudhury and Soni [44] provided a generic framework design for sensor-based HAR while analyzing cutting-edge methods and the challenges encountered by them. Imanzadeh et al. [45] collected a new dataset using gyroscopes, magnetometers, and accelerometers for sensor-based HAR. Thakur et al. [46] applied the grid search technique for hyperparameter tuning in a twin SVM framework. To reduce the training time and number of trainable parameters, Boudjema et al. [47] used cascaded training procedure with a CNN-based framework for HAR.

Gholamrezaii and AlModarresi [48] argued that eliminating pooling layers from 1D and 2D CNNs and incorporating strides in the convolutional layers could help reduce computational time without compromising the performance on sensor datasets. Geravesh and Rupapara [49] proposed a novel DS-MLP (Deep Stacked-Multi Layer Perceptron) model that employed MLP as a base learner while using an artificial neural network as a meta-learner. Hasan et al. [50] highlighted the positive influence of time series data augmentation on sensor-based HAR. Acknowledging that recurrent architectures for HAR may prove to be computationally burdensome, especially during deployment on embedded devices, Raja Sekaran et al. [51] devised a lightweight architecture by utilizing parallelly arranged convolutional heads. Based on the premise that random forests extract more interpretable features when compared with other methods, Thakur and Biswas [52] used a modified random forest for optimal feature selection from sensor data. Ghalan and Aggarwal [53] developed a novel stacked CNN framework to deal with semantically comparable daily activities. Ouyang and Pan [54] came up with a weighted voting model for HAR using single-axis and multi-axes correlation data extracted from wearable sensors. Zhang et al. [55] proposed a hybrid CNN-Transformer model to extract local and global information from sensor data. To forecast how much better a given HAR model will perform after retraining with more data, Tanigaki et al. [56] introduced a novel approach that does this before any new data is collected. Dua et al. [57] integrated the Inception CNN with GRU to capture both spatial and temporal features from sensor data. Ankalaki and Thippeswamy [58] proposed a novel activation function that overcomes the dying gradient problem faced by ReLu variants and demonstrated its ability to capture improved feature representation by using it in the hidden layer of a 1D CNN network. To overcome the problems posed by weakly annotated data and data distribution differences, Prabono et al. [59] proposed a multiple-instance domain adaptation approach for sensor-based HAR. Mhalla and Favreau [60] employed a particle filter theory-based domain adaptation approach to develop personalized HAR classifiers. Zebhi [61] used continuous wavelet transform and short-time Fourier transform to model the problem of sensor-based HAR as one of 3D volume classification. Lattanzi et al. [62] introduced a lightweight trigger to address the significant issue of limited battery life in sensor-based HAR. Liang et al. [63] presented a novel 3D attention module capable of incorporating both spatial and weight information of each channel in sensor data. Pareek et al. [64] eliminated the temporal global average pooling layer from the 3D CNN architecture and used an attention layer to capture temporal information for activity recognition. Bodhe et al. [65] proposed a hybrid model for HAR comprising CNN for extracting spatial

information, fuzzy C-means clustering for data reduction, and LSTM for capturing temporal dependencies. In [66], Stuchbury-Wass et al. investigated the suitability of multimodal attention networks for HAR using earables, i.e., wearable sensor devices for the ears.

2.2 Vision-based Human activity recognition

Technological advancements have made RGB and RGB-D cameras more accessible and affordable. Vision-based HAR systems aid surveillance for large regions. Rangasamy et al. [67] proposed a VGG-16-based transfer-learning framework for hockey activity recognition. When studying the use of LSTM in vision-based activity recognition, in an initial work, Baccouche et al. [68] used an LSTM-RNN approach to classify soccer actions of the MICC-Soccer-Actions-4 dataset using both visual and motion contents. Grushin et al. [69] demonstrated how LSTMs maintain classification performance even when subjected to unfavourable experimental conditions and how they outperform temporal pooling by taking into account the temporal ordering of video frames. In [70], Veeriah et al. proposed a differential RNN wherein handcrafted features concatenated from successive video frames were fed into the LSTM network.

Donahue et al. [71] proposed the long recurrent convolutional network, a hybrid model that brought together CNN and LSTM for visual feature learning and temporal recursion, respectively. Carreira and Zisserman [72], Ng et al. [73] and Qiu et al. [74] investigated the use of LSTMs in combination with CNNs for activity recognition amidst other prominent deep learning methods, including 3D CNN, a two-stream network, a two-stream inflated 3D CNN, and a 3D-fused two-stream network. Hybrid networks built using LSTM on top of CNNs have shown promising results in capturing spatio-temporal features. Núñez et al. [75] used a combination of CNN and LSTM for skeleton-based human activity recognition. Pandya et al. [76] obtained an accuracy of 94.20% on the UCF-101 dataset [77] using an LRCN model using the InceptionV3 pre-trained network. Negi and Kumar [78] employed ResNet-50 to perform vision-based HAR on the UTKinect dataset via transfer learning. Nguyen and Ribeiro [79] integrated CNNs with cutting-edge sequence learning models such as RNNs and transformers to recognize human activities from video sequences. To further boost the accuracy of video-based HAR, Ge et al. [80] proposed an attention-integrated CNN-LSTM architecture.

Vishwakarma and Dhiman [81] proposed a robust framework for HAR by combining Difference of Gaussian (DoG)-based spatiotemporal interest points (STIP) and Spatial Distribution of Gradients (SDGs). Nasir et al. introduced a real-time HAR model, HAR-EDNet [82], that can accurately identify human actions in unconstrained settings. Dharmalingam and Palanisamy [83] extracted and augmented features of location, shape, and speed of a person to formulate a novel feature descriptor for HAR. Shrivastava et al. [84] extracted interaction features to understand the role of each individual in multi-person activities. To reduce the computational complexity of CNN-based architectures for HAR, Kushwaha et al. [85] explored the impact of information fusion at intermediate network layers. Zhao et al. [86] encoded the 3D coordinate data of human body nodes into colour images for activity recognition using an improved CNN. By combining object appearance and motion information via feature fusion, Kushwaha et al. [87] developed a view-invariant feature descriptor for HAR. Vishwakarma and Jain [88] developed a computationally efficient framework for real-time 3D HAR using skeleton and point-based methods. Riahi et al. [89] proposed to utilize key frames from video sequences to generate dynamic images for HAR. Samma and Sama [90] optimized the YOLO computer vision system for HAR in

drone images using a two-layer particle swarm optimizer. Stressing the need to select representative frames from a given video instead of utilizing the entire frame sequence, Alfasly et al. [91] proposed a novel two-stage approach for HAR involving discriminative and representative frame selection. Varshney and Bakariya [92] designed a two-stream network for HAR, comprising a spatial stream based on an ImageNet-pretrained network and a temporal stream based on enhanced motion history images. Sun et al. [93] merged GAN and LSTM networks to accurately identify tennis activities. Sinha and Kumar [94] introduced an improved segmentation technique to extract features such as bag of visual words, local gradient threshold pattern, and grey level co-occurrence matrix for vision-based HAR. Kushwaha et al. [95] designed a computationally efficient CNN-based architecture to capture the complex and varied visual patterns in activity data. The authors stacked together small micro-networks, each catering to a distinct receptive field, to consider multi-scale features. Singh et al. [96] proposed a four-stream network for HAR by utilizing dynamic images and depth images from front, side, and top views.

3 Research gaps, challenges and key ideas

According to recent surveys [97], most HAR studies are based on sensor-based data instead of other types of data, including videos, Wi-Fi or RFID data. This can be attributed to the availability of sophisticated smartphone devices with embedded sensors and minimal privacy intrusion in capturing data. However, in light of the challenges posed by sensor-based HAR, including high costs, setup issues, and power consumption, it is necessary to develop improved techniques capable of accurately identifying human activities directly from video data. Video-based HAR studies also encounter several significant challenges [17]. A single activity might be performed in several different ways by people with varied body proportions. The angle or perspective with which the subject performing a particular activity is viewed also gives rise to multiple variations. The subject's appearance and motion patterns can vary considerable across different possible views. The environment in which videos are recorded for activity recognition also has a significant bearing on the performance of a framework. While most HAR frameworks developed in recent times work well in indoor environments, their performance degrades rapidly in cluttered and dynamic backgrounds of outdoor environments. The same subject might perform a given activity differently at different times. This type of intra-class variability caused by anthropometric disparities can complicate the framework design for a single activity. In some cases, several activities may be performed simultaneously, like brushing one's teeth while walking or talking on the phone while laying down. This further complicates both sensor and vision-based activity recognition. Collaborative activities involving multiple subjects interacting with each other constitute another challenging case. further boost the accuracy of video-based HAR, Ge et al. [80] proposed an attention-integrated CNN-LSTM architecture.

This study focuses on developing a high-performing ensemble model for video-based HAR. Our experimental findings demonstrate how the weighted voting ensemble approach can significantly boost the performance of a given architecture without the need for complex modifications. Combining the capabilities of the best-performing transfer-learning-based hybrid models helps achieve promising results with a simple and straightforward framework. To evaluate their approach, several studies based on vision-based HAR still use the UCF-101, Weizmann and KTH datasets. In order to simulate more research on similar

datasets, this paper demonstrates its findings on videos from the relatively less-explored STAIR Actions dataset.

4 Architecture for human activity recognition

This section seeks to elaborate upon the proposed ensemble architecture and the hybrid models constituting it.

4.1 Problem statement

Let us consider V as a representation of a visual sample space. With respect to vision-based human activity recognition, each dataset sample consists of a video. Further, each example is associated with a class label denoted as C^i. Simply put, each instance comprises two components: a video, and a classification label indicating the activity being performed in that video. The dataset can be represented as shown in Eq. (1).

$$H = \{(V^0, C^0), (V^1, C^1), \ldots, (V^i, C^i), \ldots, (V^{n-1}, C^{n-1})\} \tag{1}$$

where, H denotes the collection of dataset pairs, V^i denotes the data corresponding to visual modality, C^i stands for the class label of the i^{th} video, and n represents the total number of dataset samples. The proposed framework is designed to acquire a mapping function $F : V \rightarrow C$ from the inputs that constitute set H. For the activity classification task, $C^i \in \{walking\ with\ stick,\ standing\ up,\ sitting\ down,\ eating\ meal,\ drinking\}$.

4.2 Proposed Model

The ensemble learning process combines multiple models to get better classification accuracy. To prepare an ensemble model, we chose the best-performing hybrid models from each category of pre-trained CNN networks used. These models were – VGG19 + LSTM, ResNet50 + LSTM, Xception + LSTM and DenseNet121 + LSTM with test set accuracies of 82.73%, 73.64%, 83.64% and 88.18% respectively. A detailed comparison of the performance parameters of these models, including test-set accuracy, amongst others, has been given in Section 4.5. The sequence of steps undertaken to achieve the proposed weighted voting ensemble model is shown in Fig. 2. Fig. 3 illustrates the architecture for this ensemble model.

Before the weighted ensemble, a simple model averaging ensemble was also tested. The result obtained from this model served as a measure to compare the results

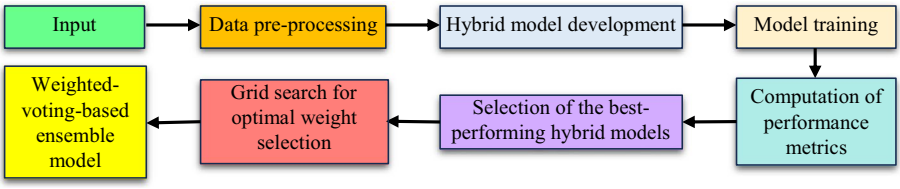


Fig. 2 Sequence of steps undertaken to devise the proposed ensemble model

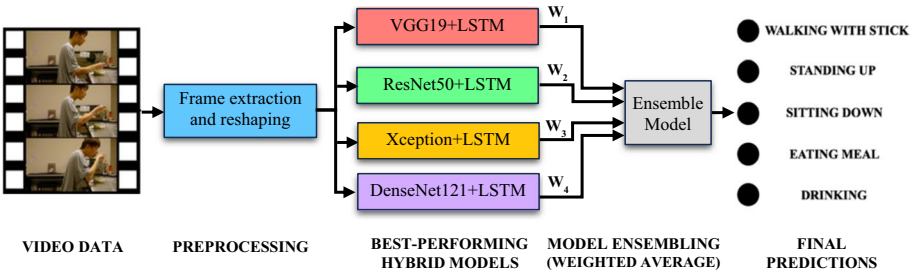


Fig. 3 Weighted voting ensemble model architecture

obtained from weighted ensembles. Under the simple model averaging ensemble, after obtaining the prediction probabilities for each activity from each of the four models, the probabilities for each class were added, and the class with the maximum value for the sum of predicted probabilities was chosen as the final prediction. In this approach, all the models have equal weights (equal to $\frac{1}{n}$ where n is the total number of models); hence, the model outputs are non-weighted.

Under the weighted average ensemble approach, the contribution of each ensemble member model is indicated by a weight value proportional to its performance. As opposed to directly adding the prediction probabilities in the simple ensemble model, a weighted sum is calculated in this method. The weighted sum is given by the value $\sum_{i=1}^{n} p_i w_i$, where n is the number of models constituting the ensemble, p_i the prediction probability as obtained from the ith model and w_i the weight assigned to it. The weight values chosen are typically between 0 and 1 and are assigned such that the sum of weights across all contributing models equals one. To reach the best possible accuracy using our trained models, we used the exhaustive grid search technique to calculate the combinations of the most optimal weights by testing all possible cases. In case the sum of the weights assigned by the grid search was not one, normalization was carried out to achieve the same. Table 14 in Section 4.5 summarizes the results obtained for different weight combinations.

4.3 Convolutional Neural Network (CNN)

CNNs are commonly employed in various image recognition applications [98, 99] due to their capacity to extract spatial characteristics from images. These networks can automatically learn and extract distinctive features from images, including those that may be challenging to characterize or quantify. CNNs are structured using a layered architecture comprising three types of layers: convolutional, pooling, and fully-connected. The multi-layered structure of CNNs allows them to capture low, mid and high-level features, thus emulating the hierarchical learning process of the human brain. Figure 4 seeks to categorize the CNNs used in this study on the basis of architectural configuration. Table 1 outlines the major contributions, strengths and limitations of the different architectures. Table 2 lists the ten different CNN models used in conjunction with LSTM and summarizes their key details.

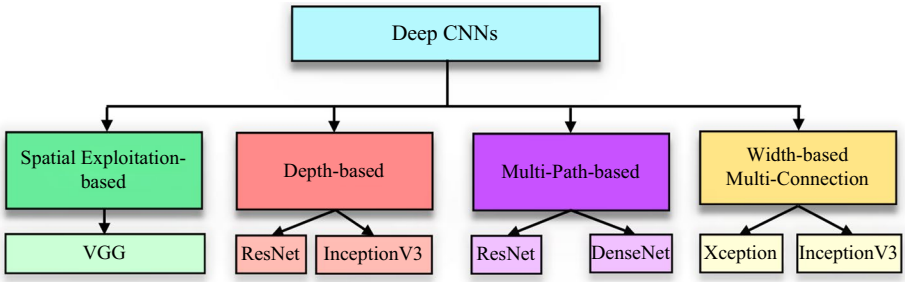


Fig. 4 Taxonomy of CNN architectures used in this study

4.4 Long short-term memory (LSTM)

LSTM has proven to be a boon to applications involving sequential relationships. It is a special type of RNN network proposed by Hochreiter and Schmidhuber [100] that utilizes memory cells to maintain state value over a long time and output information. Its architecture comprises memory blocks, with each block holding a number of memory cells. As shown in Fig. 5, a memory cell consists of three non-linear gates that control the flow of signals into and outside the memory cells. These gates are – the input gate to decide what new information is to be stored in each cell, the forget gate to decide whether to retain or forget certain information and the output gate to choose useful information from the current cell state to display as output. Mathematically, the functioning of an LSTM unit can be expressed as shown in Eq. (2), Eq. (3), Eq. (4), Eq. (5), and Eq. (6).

$$i_t = \sigma\left(W_i \otimes \left[h_{t-1}, x_t\right] + b_i\right) \tag{2}$$

$$f_t = \sigma\left(W_f \otimes \left[h_{t-1}, x_t\right] + b_f\right) \tag{3}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes tanh\left(W_c \otimes \left[h_{t-1}, x_t\right] + b_g\right) \tag{4}$$

$$o_t = \sigma\left(W_o \otimes \left[h_{t-1}, x_t\right] + b_o\right) \tag{5}$$

$$h_t = o_t \otimes tanh\left(c_t\right) \tag{6}$$

where i_t, f_t, c_t, o_t, h_t, and x_t represent the input gate, forget gate, memory unit, output gate, hidden state, and input at time t respectively. W and b denote the response weight and bias respectively. *tanh* and σ stand for the hyperbolic tangent and sigmoid activation functions respectively while $\otimes$ symbolizes the multiplication operation of elements.

4.5 Hybrid CNN + LSTM architecture

This research work uses a combination of CNN and LSTM for 2D video human activity recognition. While CNNs alone do not take into account the temporal dependencies of the

Table 1 Summary of the different CNN architectures used to build the hybrid CNN-LSTM networks

Network	Major Contribution	Strengths	Limitations
VGG	Homogenous architecture with small kernels	**1.** Increased number of layers with smaller kernel size. **2.** Increased non-linearity.	**1.** High computational cost of fc layers. **2.** Vanishing gradient
ResNet	Residual learning and skip connections	**1.** Helped combat the vanishing gradient issue. **2.** Residual connections enable faster training and better convergence.	**1.** Complex architecture. **2.** Training can be computationally expensive.
InceptionV3	Replaced large-sized filters with smaller ones and tackled the issue of representational bottleneck.	**1.** Utilized asymmetric filter combinations and bottleneck layers to reduce the computational demands. **2.** Filters of different sizes aid in capturing a wide range of features in high-resolution images.	**1.** Increase in space and time complexity. **2.** Complex architecture. **3.** Lack of homogeneity.
Xception	Depth-wise separable convolutions	**1.** Introduced depth-wise separable convolutions. **2.** Learns good abstractions via cardinality.	**1.** Increased computational costs.
DenseNet	Cross-layer information flow	**1.** Allows maximum data flow across network layers. **2.** Eliminates redundancy of re-learning feature maps.	**1.** Computationally expensive. **2.** Memory-intensive.

Table 2 CNN architecture summary (https://keras.io/api/applications/)

Network	Depth	Size (in MB)	Parameters (in millions)	Top-1 Accuracy on ImageNet dataset (%)
VGG16	16	528	138.4	71.3
VGG19	19	549	143.7	71.3
ResNet50	107	98	25.6	74.9
ResNet101	209	171	44.7	76.4
ResNet152	311	232	60.4	76.6
InceptionV3	189	92	23.9	77.9
Xception	81	88	22.9	79
DenseNet121	242	33	8.1	75
DenseNet169	338	57	14.3	76.2
DenseNet201	402	80	20.2	77.3

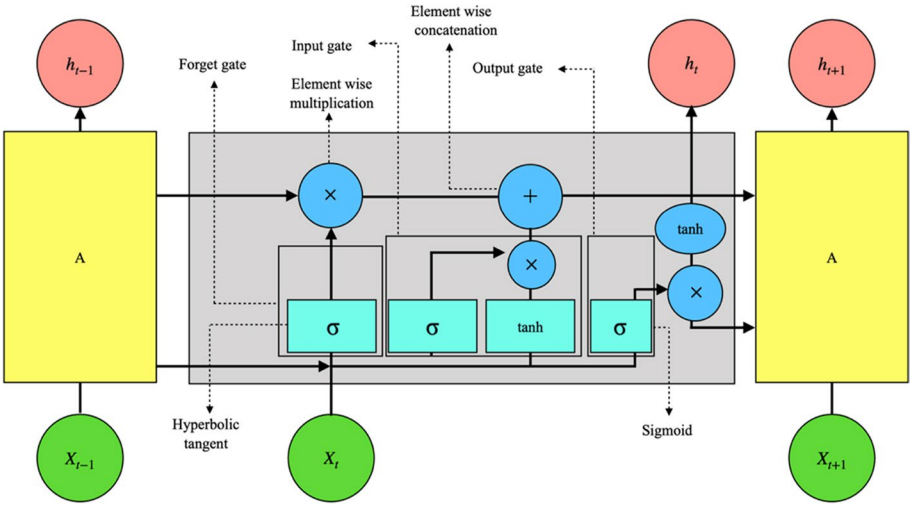


Fig. 5 Structure of LSTM unit. Here, X is the input of the memory, and h is the prediction at time t

different frames of a video, LSTMs, a special type of RNN, can make up for this shortcoming by maintaining observations in memory for extended periods using their gating mechanism.

For this study, ten CNNs have been used to process individual video frames. The LSTM network helps exploit the variations between frames to achieve more accurate predictions. As stated earlier, the LSTM network uses memory cells to store and output information. This leads to the better discovery of long-range temporal relationships and helps overcome the gradient vanishing and exploding issues encountered in training RNN networks. Twenty video frames with equivalent distance were extracted from each input video for feature extraction and temporal relationship classification. In the hybrid CNN+LSTM architecture, data transmission is efficient only when the data format remains consistent at the network junctions. The feature dimension of the LSTM

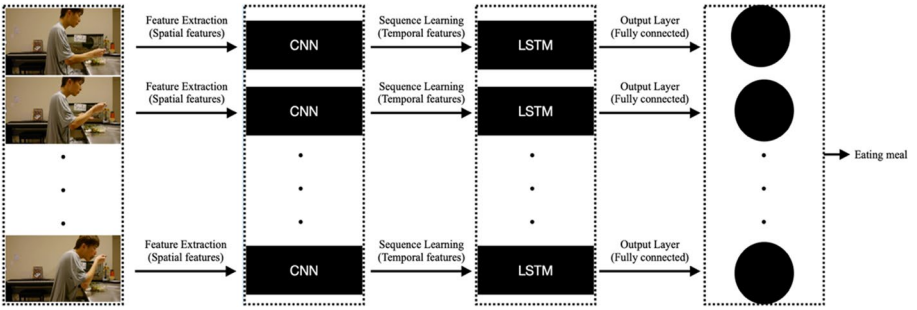


Fig. 6 Overall framework of the combination of CNN with LSTM. The extracted spatial features from the CNN pass through the LSTM to create the temporal features

model should be determined by the output of the CNN model to maintain consistency in feature extraction. Figure 6 depicts the core sequence of steps for the approach used. The reconfiguration of standard CNNs first entailed deleting the fully connected layers while keeping the convolutional and pooling layers as the core network. To regulate the number of output features, a global average pooling layer was used to obtain the CNN output. The sequential frames are encoded by the CNN model to generate a series of feature vectors, which are then supplied as input to an LSTM layer with 32 units.

4.6 Transfer Learning

For this work, we used different networks pre-trained on the standard ImageNet [101] dataset for feature extraction from input video frames by eliminating the top layer used for classification. Ten state-of-the-art deep learning models - VGG16, VGG19 [102], ResNet50, ResNet101, ResNet152 [103], Inception V3 [104], Xception [105], DenseNet121, DenseNet169 and DenseNet201 [106] were used as the pre-trained models for feature extraction. The output of the base CNN model passes through a global average pooling layer followed by a time-distributed layer wrapping to apply the feature extraction layer to each temporal slice or frame of input. This is followed by the LSTM to identify temporal features within the frame sequence. The combined features are then passed to a fully connected layer to predict a single activity for the given input sequence, as visualized in Fig. 7.

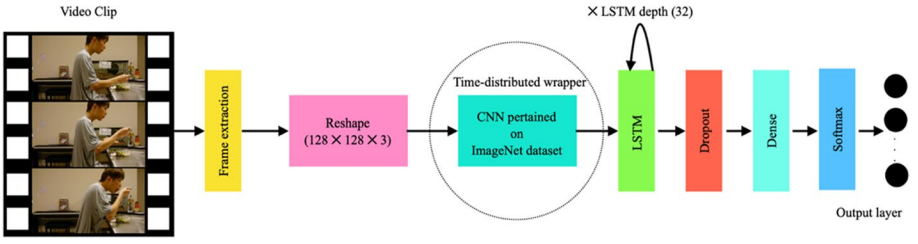


Fig. 7 Model architecture

Fig. 8 Sample frames for the five activities – eating meal, drinking, standing up, walking with stick, sitting down

Table 3 Number of videos for each activity chosen for activity recognition

Class/Activity	Drinking	Eating meal	Sitting down	Standing up	Walking with stick
Number of videos in dataset	408	441	886	503	923

5 Experimental Settings

This section describes the experimental settings and the methodology followed to establish the proficiency of the proposed framework.

5.1 Dataset Description

We used videos from the STAIR (Software Technology and Artificial Intelligence Research Laboratory) Actions dataset [107], a large-scale video dataset for activity recognition and classification. The dataset consists of trimmed video clips of 5-6 seconds each for everyday human activities, with each video clip being associated with a single label. The dataset video clips have either been procured from YouTube or prepared by crowdsourcing workers. For this paper, five classes of activities were chosen: walking with a stick, standing up, sitting down (from solo action set of activities), eating a meal and drinking (from kitchen related set of activities). Sample frames for these activities have been depicted in Fig. 8. The number of videos per class is shown in Table 3.

5.2 Performance Metrics

To evaluate the performance of the models, the metrics adopted include AUROC, AUPRC, precision, specificity, sensitivity (or recall) and overall accuracy. Section 4.5 discusses the need for these metrics and compares the values obtained for the different models.

The metrics are defined in Table 4, where TP, TN, FP and FN stand for true positives, true negatives, false positives and false negatives, respectively.

Table 4 Performance metrics

Performance metric	Description and Mathematical Formula
Precision	Precision, or positive predictive value, identifies the proportion of positive predictions that were actually true. It is the ratio of the true positives to the total positives predicted. $Precision = \frac{TP}{TP+FP}$
Specificity	Specificity, or true negative rate, identifies the proportion of true negatives that were correctly predicted by the model. It is the ratio of true negatives to all the negatives in the ground truth. $Specificity = \frac{TN}{TN+FP}$
Sensitivity	Sensitivity, or hit rate or true positive rate, identifies the proportion of true positives that were correctly predicted by the model. It is the ratio of true positives to all the positives in the ground truth. $Sensitivity = \frac{TP}{TP+FN}$
Accuracy	Accuracy is the measure of how often the classifier is correct. It is the ratio of the number of correct predictions to the total number of predictions. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
AUROC (Area under the ROC curve)	This is a metric that defines the degree to which the model is able to discriminate or separate the different classes. The higher the value, the better the model is at distinguishing between the different classes. The ROC curve is a TPR vs FPR curve with TPR on the y-axis and FPR on the x-axis. Since ours is a multi-class problem, we have used OvR (One vs Rest) method to plot the curves.
AUPRC (Area under the precision-recall curve)	This is a useful metric to evaluate output quality of a model for an imbalanced dataset. The higher the value, the higher the precision (low FPR) and higher the recall (low FNR). The precision-recall curve is a precision vs recall curve for different thresholds with precision on the y-axis and recall on the x-axis.

5.3 Video frame extraction

Video-based HAR requires us to extract frames to study the temporal relationships between them. The OpenCV library was used to extract 20 equidistant frames of dimension 128×128 for each video in the dataset. Figure 9 illustrates the frames extracted for a video of the class – 'eating meal'.

5.4 Hyperparameters

Ten different CNN + LSTM network models formed using ten different architectures, namely VGG16, VGG19, ResNet50, ResNet101, ResNet152, InceptionV3, Xception, DenseNet121, DenseNet169 and DenseNet201 were trained on the videos for five classes from the dataset. In order to determine and establish the best possible performance for the different models compared, we trained several models by experimenting with different values for frame dimensions, batch size, epochs, patience value for early stopping call-back and learning rate. All models have been compared, keeping the frame height and width 128, batch size 4, epochs 50, early stopping call-back with a patience value of 10 and using the Adam optimiser with cross-entropy categorical loss calculation [108]. Early stopping call-back based on validation loss helped prevent overfitting of the models on the training dataset so as to ensure that they are able to perform well on unseen inputs. The training process terminated when there was no improvement in performance for 10 consecutive

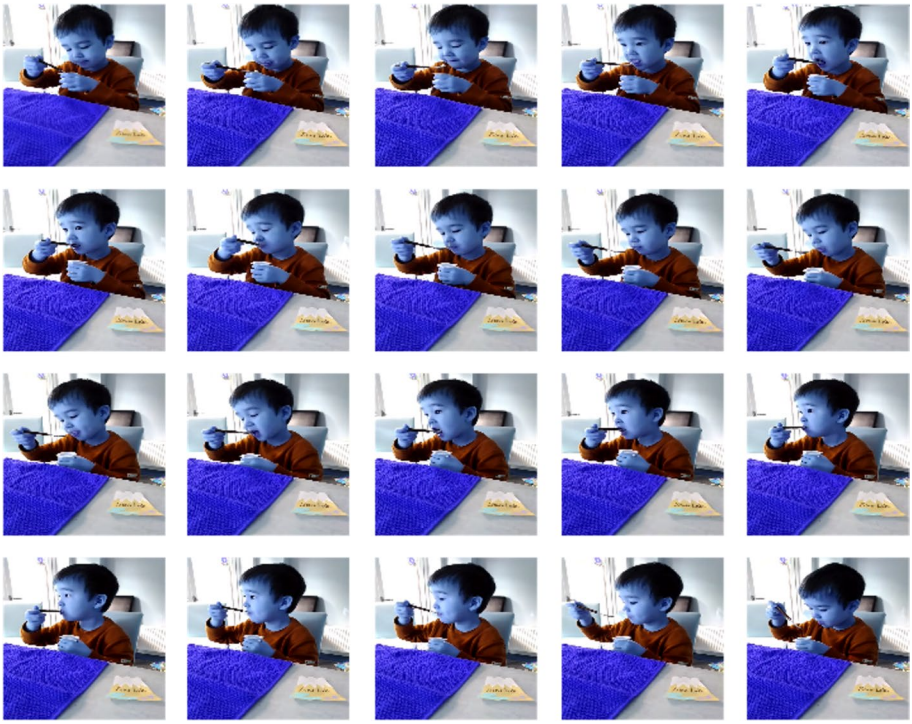


Fig. 9 Video frames extracted from a sample video from the class – 'eating meal'

Table 5 Number of parameters for each of the network architectures

Network	Total parameters	Trainable parameters	Non-trainable parameters
VGG16 + LSTM	14,785,669	70,981	14,714,688
VGG19 + LSTM	20,095,365	70,981	20,024,384
ResNet50 + LSTM	23,855,301	267,589	23,587,712
ResNet101 + LSTM	42,925,765	267,589	42,658,176
ResNet152 + LSTM	58,638,533	267,589	58,370,944
InceptionV3 + LSTM	22,070,373	267,589	21,802,784
Xception + LSTM	21,129,069	267,589	20,861,480
DenseNet121 + LSTM	7,174,021	136,517	7,037,504
DenseNet169 + LSTM	12,861,317	218,437	12,642,880
DenseNet201 + LSTM	18,573,189	251,205	18,321,984

epochs. The number of trainable and non-trainable parameters for each of the ten hybrid models has been stated in Table 5.

5.5 Implementation details

The models in this study were trained and tested using Google COLAB Pro+ equipped with NVIDIA V100 GPU, 52 GB of system RAM, 150 GB of hard disk space, and 16 GB of GPU RAM. The models were constituted and trained using Python programming language with Keras libraries.

6 Results and discussion

Under this section, we first present a review of the individual models based on the performance metrics defined in Section 4.2, following which the results of the ensemble model have been discussed.

6.1 Performance of the individual hybrid models

This section compares the results of the ten hybrid models on the five classes of activities chosen from the STAIR Actions dataset.

The CNN-LSTM models were tested on a set of 230 videos spread over the five classes. As depicted in Fig. 10, the dataset showed class imbalance (walking with a stick 29.2%, standing up 15.9%, sitting down 28.03%, eating meal 13.95% and drinking 12.9%). Thus, we used confusion matrices and graph plots of receiver operation characteristics (ROC) and precision-recall curves for each class to evaluate, analyze, and compare the trained models. Both AUROC and AUPRC values were calculated for each class for each model. The performance metrics calculated for each class include specificity, sensitivity, and

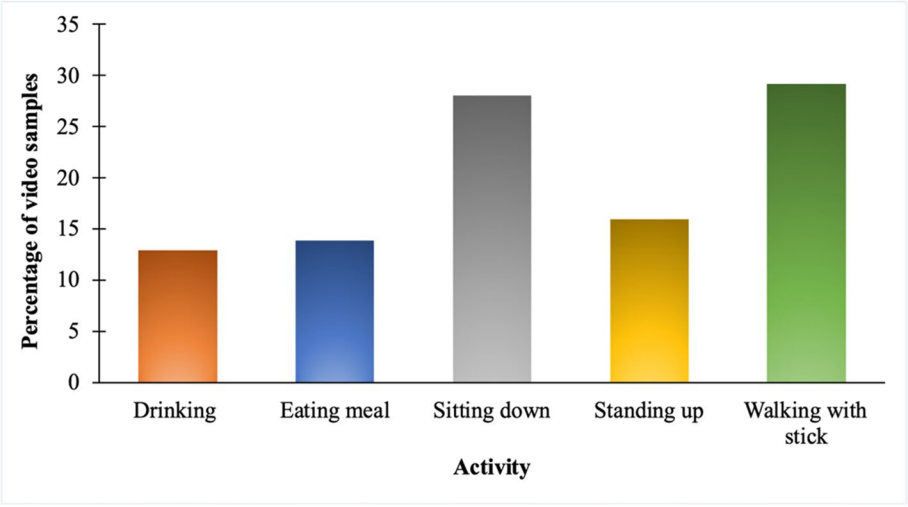


Fig. 10 Class imbalance in dataset

precision. After experimenting with different values for batch size, batches of 4 gave optimum results, so all models were compared using a batch size of 4.

The accuracy and loss values during training, validation and testing have been shown in Table 6. While comparing the number of epochs taken by each model to achieve the minimum value for validation loss, we observed that DenseNet169, DenseNet121 and Xception reached the lowest value at just 9, 12 and 9 epochs, respectively. Within these epochs, they reached accuracy values above 90%. When studying training parameters, it was found that five of the best-performing models, i.e., VGG16, Xception, DenseNet121, DenseNet169 and DenseNet201, were able to achieve a training accuracy of 100%. The model trained using Xception achieved the lowest value for training loss, followed by DenseNet121, DenseNet169 and DenseNet201. From this data, it can be inferred that all variants of DenseNet architecture exhibited the best performance for the task at hand. The confusion matrices shown in Fig. 11 display the number of video samples correctly and incorrectly identified by each model for each class.

Learning curves are graphical representations that depict the evolution of learning performance over time. They serve as valuable tools for evaluating the fitness of a model. Table 7 displays the curves of training and validation loss and accuracy as the number of epochs increases. The training and validation loss curves exhibit a consistent and gradual decrease, leading to the convergence of loss values at a stable point. The validation accuracy curves exhibit a steady increase as the number of epochs progresses. The learning curves impart empirical support for the adequacy of the models' fit and effectiveness of their training. Table 8, Table 9, Table 10 and Table 11 summarize the precision, sensitivity, specificity, and accuracy values.

When studying the individual findings for each model, the highest AUROC of walking with stick was achieved by DenseNet201 (1.0000), DenseNet121 (0.9994), DenseNet169 (0.9992) and VGG16 (0.9996). For recognition of standing up, the best models were DenseNet169 (0.9689) and DenseNet201 (0.9665). For sitting down, the highest AUROC values were obtained using DenseNet169 (0.9825) and DenseNet121 (0.9782). The activity of eating meal was most confidently predicted by DenseNet201 (0.9962) and DenseNet169 (0.9834). Finally, for drinking, the the models with best results were Xception (0.9862), DenseNet121 (0.9859) and DenseNet201 (0.9852). Figure 12 presents the ROC curves for each class and model, while Table 12 summarizes the AUROC values. The highest values of AUPRC for walking with stick were achieved by DenseNet201 (1.0000), DenseNet121 (0.9989) and DenseNet169 (0.9986) and for standing up by ResNet101 (0.9038). The best classification for sitting down was obtained using DenseNet169 (0.9425). DenseNet201 (0.9829) helped achieve the highest AUPRC for the activity of eating meal. Finally, for drinking, the models with the best results were Xception (0.9624) and DenseNet121 (0.9614). Figure 13 presents the precision-recall curves for each class and model, while Table 13 summarizes the AUROC values.

6.2 Ablation study

An ablation study aims to determine the essentiality of each component in a given architecture to validate its inclusion. This process is usually carried out by methodically removing each component and analyzing the effect of its absence on the overall model performance. This study utilizes a hybrid CNN-LSTM model to collect both spatial and temporal features for HAR. We conduct an ablation study to determine the efficacy of LSTM in this hybrid model. Figure 14 illustrates the outcomes achieved with and

Table 6 Loss and accuracy during training, validation and testing

Network	Best epoch	Train loss	Train accuracy	Validation loss	Validation accuracy	Test loss	Test accuracy
VGG16 + LSTM	29	0.0061	1.0000	0.3264	0.8939	0.5011	0.8227
VGG19 + LSTM	17	0.1736	0.9413	0.3575	0.8864	0.5363	0.8273
ResNet50 + LSTM	43	0.4829	0.8314	0.7869	0.7727	0.8202	0.7364
ResNet101 + LSTM	21	0.3808	0.8826	0.7080	0.7879	0.7714	0.7273
ResNet152 + LSTM	37	0.2612	0.9242	0.8216	0.7727	1.0443	0.6455
InceptionV3 + LSTM	17	0.0337	0.9962	0.4957	0.8561	0.7321	0.7182
Xception + LSTM	12	0.0012	1.0000	0.2341	0.9167	0.5467	0.8364
DenseNet121 + LSTM	9	0.0031	1.0000	0.3778	0.8712	0.3452	0.8818
DenseNet169 + LSTM	9	0.0035	1.0000	0.2443	0.9242	0.3467	0.8682
DenseNet201 + LSTM	15	0.0048	1.0000	0.2838	0.9091	0.3523	0.8682

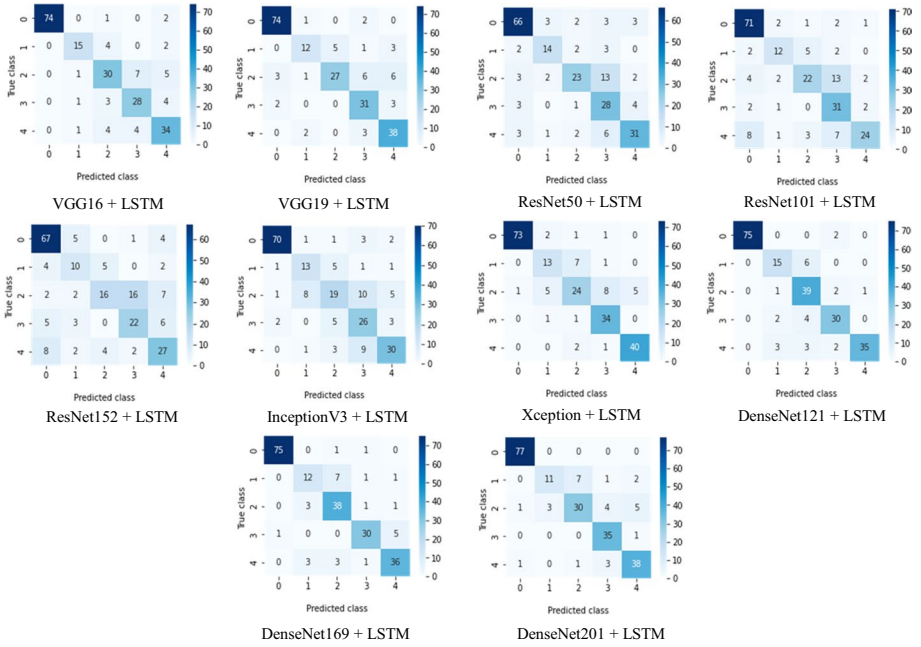


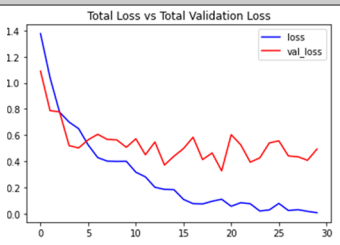
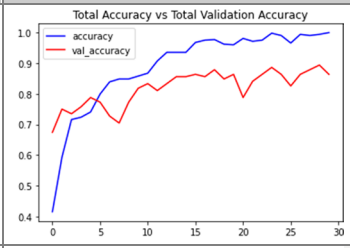
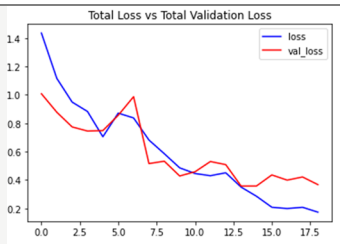
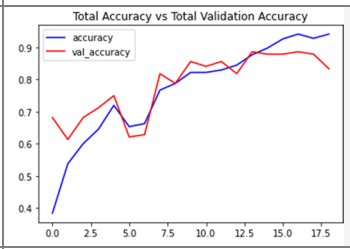
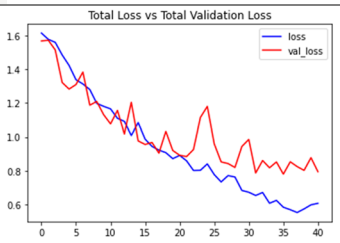
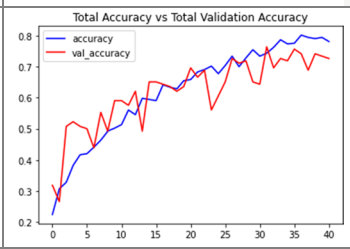
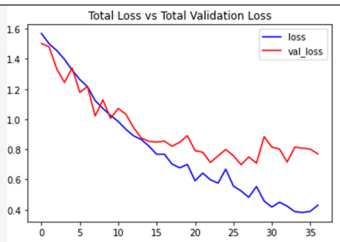
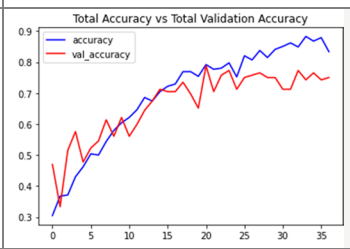
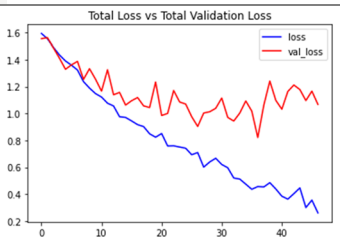
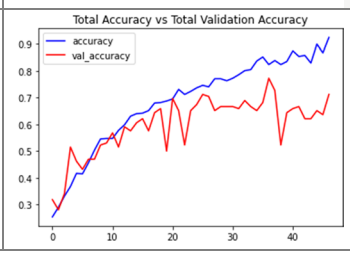
Fig. 11 Confusion matrices for all 10 models. The class number and corresponding labels are, 0 – walking with stick, 1 – standing up, 2 – sitting down, 3 – eating meal and 4 – drinking

without LSTM for each of the ten CNNs utilized in this study. There is a striking contrast in performance between the two scenarios depicted. Eliminating LSTM results in a notable decrease in test-set accuracy, demonstrating its capacity to extract the necessary temporal information for activity recognition. The CNN can only capture the spatial features on the frame level and does not account for the sequential relationships across frames. Since activities such as walking, sitting, standing, eating, and drinking are constituted by characteristic sequences of actions undertaken by a subject across successive video frames, it is imperative to employ a sequential learning model to recognize them accurately.

6.3 Weighted voting ensemble model

When a finite number of models are trained to perform the same task, ensemble learning is used to merge multiple models to improve performance and reduce instability and variance. The application of the weighted voting ensemble methodology (Fig. 15) outperformed the results of all individual hybrid models compared. The test set accuracy for the simple averaging ensemble model was 0.8954, a slight increase from the best performing individual model. Table 14 summarizes the best weight combinations along with their respective test set accuracies for the proposed model. The highest test set accuracy of 0.9318 was achieved using the set of weights {0.1333, 0.1333, 0.2667, 0.4667} for models ResNet50 + LSTM, VGG19 + LSTM, Xception + LSTM and DenseNet121 + LSTM respectively.

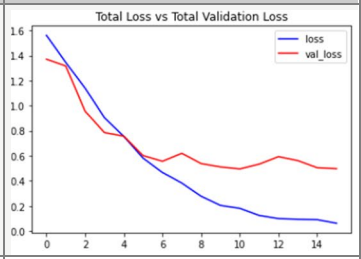
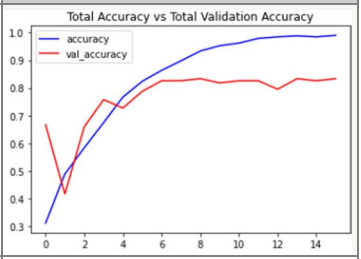
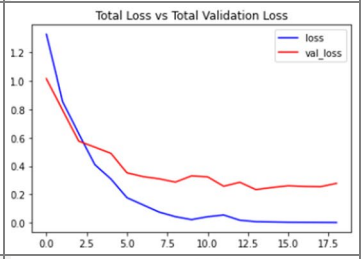
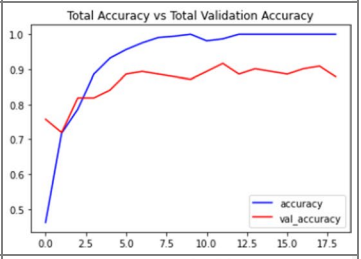
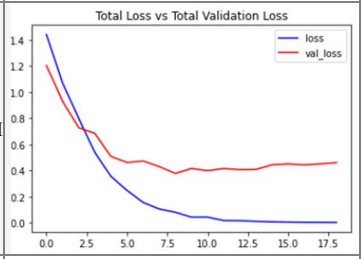
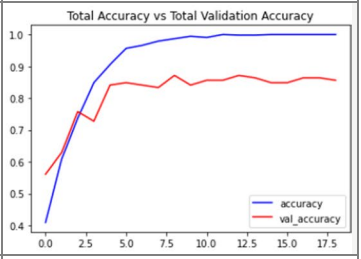
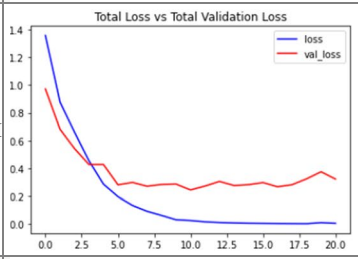
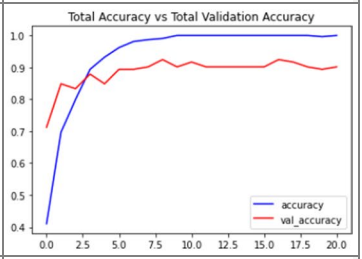
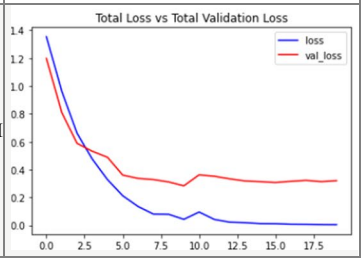
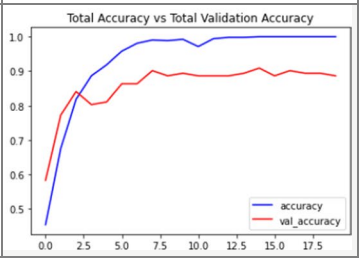
Table 7 Loss and accuracy curves for each of the network architectures

Network	Loss Curves	Accuracy Curves
VGG16 + LSTM		
VGG19 + LSTM		
ResNet50 + LSTM		
ResNet101 + LSTM		
ResNet152 + LSTM		

6.4 Discussion of results achieved using individual hybrid models and weighted voting ensemble model

This section serves to analyze the results achieved using the individual hybrid models and the weighted voting ensemble model and understand the underlying factors that affect

Table 7 (continued)

Network	Loss Curves	Accuracy Curves
InceptionV3 + LSTM		
Xception + LSTM		
DenseNet121 + LSTM		
DenseNet169 + LSTM		
DenseNet201 + LSTM		

them. In addition, we also discuss the impact of the dataset on the results obtained and how the dataset's magnitude can affect the attained outcomes. To thoroughly evaluate and compare the performance of the ten hybrid models, a host of metrics (Table 4) were calculated,

Table 8 Precision values for each of the network architectures

Network	Walking with stick	Standing up	Sitting down	Eating meal	Drinking
VGG16 + LSTM	1.0000	0.8333	0.7143	0.7179	0.7234
VGG19 + LSTM	0.9367	0.7500	0.8438	0.72093	0.7600
ResNet50 + LSTM	0.8571	0.7000	0.7667	0.5283	0.7750
ResNet101 + LSTM	0.8161	0.6667	0.7097	0.5634	0.8276
ResNet152 + LSTM	0.7791	0.4545	0.6400	0.5366	0.5870
InceptionV3 + LSTM	0.9459	0.5652	0.5758	0.5306	0.7317
Xception + LSTM	0.9865	0.6190	0.6857	0.7556	0.8889
DenseNet121 + LSTM	1.0000	0.7143	0.7500	0.8333	0.9722
DenseNet169 + LSTM	0.9868	0.6667	0.7755	0.8824	0.8372
DenseNet201 + LSTM	0.9747	0.7857	0.7895	0.8140	0.8261

Table 9 Sensitivity values for each of the network architectures

Network	Walking with stick	Standing up	Sitting down	Eating meal	Drinking
VGG16 + LSTM	0.9610	0.7143	0.6977	0.7778	0.7907
VGG19 + LSTM	0.9610	0.5714	0.6279	0.8611	0.8837
ResNet50 + LSTM	0.8571	0.6667	0.5345	0.7778	0.7209
ResNet101 + LSTM	0.9221	0.5714	0.5116	0.8611	0.5581
ResNet152 + LSTM	0.8701	0.4762	0.3721	0.6111	0.6279
InceptionV3 + LSTM	0.9091	0.6190	0.4419	0.7222	0.6977
Xception + LSTM	0.9481	0.6190	0.5581	0.9444	0.9302
DenseNet121 + LSTM	0.9740	0.7143	0.9070	0.8333	0.8140
DenseNet169 + LSTM	0.9740	0.5714	0.8837	0.8333	0.8372
DenseNet201 + LSTM	1.0000	0.5238	0.6977	0.9722	0.8837

Table 10 Specificity values for each of the network architectures

Network	Walking with stick	Standing up	Sitting down	Eating meal	Drinking
VGG16 + LSTM	1.0000	0.9849	0.9322	0.9402	0.9266
VGG19 + LSTM	0.9650	0.9799	0.9718	0.9348	0.9322
ResNet50 + LSTM	0.9231	0.9698	0.9605	0.8641	0.9492
ResNet101 + LSTM	0.8881	0.9698	0.9492	0.8696	0.9718
ResNet152 + LSTM	0.8671	0.9397	0.9492	0.8967	0.8927
InceptionV3 + LSTM	0.9720	0.9497	0.9209	0.8750	0.9379
Xception + LSTM	0.9930	0.9598	0.9379	0.9402	0.9718
DenseNet121 + LSTM	1.0000	0.9698	0.9266	0.9674	0.9944
DenseNet169 + LSTM	0.9930	0.9698	0.9379	0.9783	0.9605
DenseNet201 + LSTM	0.9860	0.9849	0.9548	0.9565	0.9548

as reported in Section 5.1. In terms of test-set accuracy, the DenseNet architecture achieved the best results. The lower trainable parameter count of this variant, as compared to the other two, points to the conclusion that a higher number of trainable parameters does not

Table 11 Classification accuracy for each of the network architectures

Network	Walking with stick	Standing up	Sitting down	Eating meal	Drinking
VGG16 + LSTM	0.9864	0.9590	0.8864	0.9136	0.9000
VGG19 + LSTM	0.9636	0.9409	0.9045	0.9227	0.9227
ResNet50 + LSTM	0.9000	0.9409	0.8772	0.8500	0.9045
ResNet101 + LSTM	0.9000	0.9318	0.8636	0.8682	0.8909
ResNet152 + LSTM	0.8682	0.8955	0.8364	0.8500	0.8409
InceptionV3 + LSTM	0.9500	0.9182	0.8272	0.8500	0.8909
Xception + LSTM	0.9773	0.9272	0.8636	0.9409	0.9636
DenseNet121 + LSTM	0.9909	0.9454	0.9227	0.9454	0.9591
DenseNet169 + LSTM	0.9864	0.9318	0.9273	0.9545	0.9364
DenseNet201 + LSTM	0.9909	0.9409	0.9045	0.9591	0.9409

necessarily guarantee a higher accuracy. The hybrid models created using the ResNet architecture recorded the worst performance in terms of test-set accuracy. The underlying principle of the ResNet architecture is to allow for training deeper CNNs via skip connections between the front and back layers. These shortcut connections facilitate gradient propagation during model training. The DenseNet architecture, too, strives to function along the same fundamental concept, but it forms dense connections between all preceding and subsequent layers, as indicated by its name. This allows DenseNet to outperform ResNet while requiring lesser computational cost and fewer parameters. The least test-set accuracy achieved using a DenseNet variant (0.8682) exceeds the best score obtained using a ResNet variant (0.7364) by a significant margin.

ResNet variants, amongst other deep learning networks, exhibit a characteristic where the test-set accuracy of the model declines as the number of layers goes beyond a specific threshold. This trend can also be verified by the results outlined in Table 6. The test-set accuracy value declines from 0.7364 to 0.7273 as we move from ResNet-50 to ResNet-101 and further reduces to 0.6455 for ResNet-152. A similar pattern can also be seen with respect to the DenseNet variants. Therefore, it is vital that the depth of the model be chosen cautiously. Unlike its predecessors, the Inception architecture employs different-sized filters and max-pooling technique to decrease the dimensionality of input data and attain more complex feature representations. Though the hybrid models constituted using InceptionV3 and Xception possess the same number of trainable parameters, the depth-wise separable convolutions in the Xception architecture help achieve significantly better classification performance.

After having trained and evaluated the individual hybrid models, we developed an ensemble model that is capable of utilizing the strengths of diverse base learners via a weighted voting scheme. By aggregating the knowledge from acquired from various base models, the ensemble model enhances the overall predictive accuracy on the test-set. The ensemble model was constructed by selecting the best-performing variant of each of the base learner architectures, to allow for the inclusion of diverse modelling abilities and perspectives. The cumulative application of the strengths of multiple base classifiers boosts predictive accuracy, as seen in Table 14. The weighted voting scheme assigns a weight to each of the base learners to integrate their predictions and impact the final model prediction. Following extensive iterations of different combinations of weights computed using grid-search mechanism, the results in Table 14 were obtained. The most optimal weight

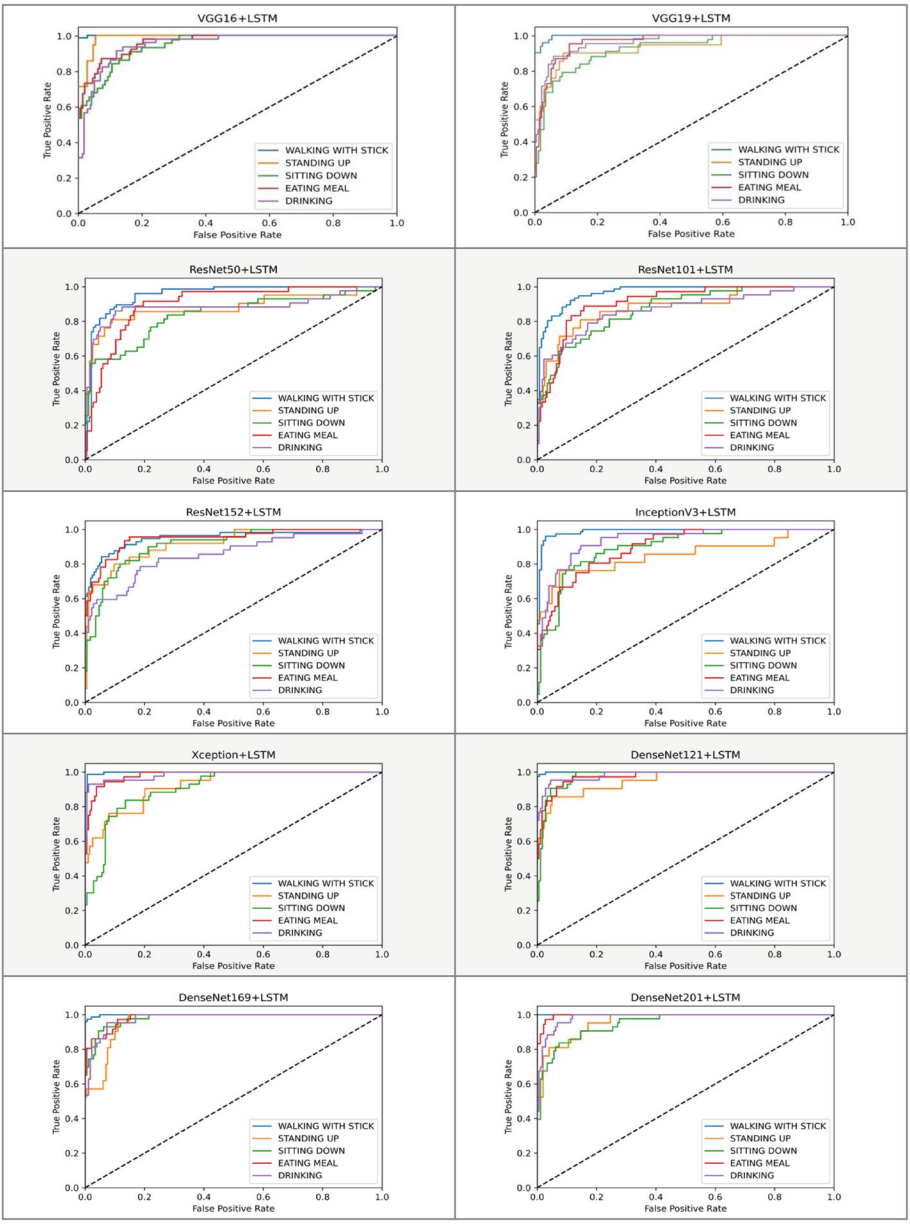


Fig. 12 ROC curves for each of the network architectures

allocation {0.1333, 0.1333, 0.2667, 0.4667} for the models ResNet50 + LSTM, VGG19 + LSTM, Xception + LSTM and DenseNet121 + LSTM respectively, helped achieve the highest test-set accuracy of 0.9318. This is a significant increase from the highest score achieved by an individual hybrid model (DenseNet121 + LSTM, 0.8818). This observation bears testament to the performance-enhancing abilities of ensembles, which are also quite

Table 12 AUROC for each of the network architectures

Network	Walking with stick	Standing up	Sitting down	Eating meal	Drinking
VGG16 + LSTM	0.9996	0.9904	0.9453	0.9610	0.9490
VGG19 + LSTM	0.9968	0.9382	0.9227	0.9598	0.9606
ResNet50 + LSTM	0.9520	0.9045	0.8534	0.8966	0.9026
ResNet101 + LSTM	0.9667	0.9038	0.8993	0.9139	0.8695
ResNet152 + LSTM	0.9399	0.8710	0.8422	0.8557	0.8651
InceptionV3 + LSTM	0.9899	0.8559	0.9042	0.8930	0.9327
Xception + LSTM	0.9986	0.9222	0.9083	0.9828	0.9862
DenseNet121 + LSTM	0.9994	0.9572	0.9782	0.9748	0.9859
DenseNet169 + LSTM	0.9992	0.9689	0.9825	0.9834	0.9773
DenseNet201 + LSTM	1.0000	0.9665	0.9482	0.9962	0.9852

intuitive since they take the predictions of diverse top-performing base learners into consideration at the same time.

The chosen data sample also has significant bearing on the results achieved. The dataset in the present scenario is of a moderate size. If a larger data sample had been chosen, with a higher number of videos as well as classes, the simple yet sturdy model under examination could yield even better outcomes. Additionally, increasing the dataset sample size also enhances the model's generalization capabilities by allowing it to learn from a wider range of scenarios and examples for each activity. While larger datasets do demand increased computational resources required for training and testing, they are employed because they are capable of capturing varied instances of activities, thus allowing for development of robust models. Increasing the number of classes will facilitate the inclusion of a practical set of activities that can be applied in daily life, thus developing a suitable system for real-world applications. As the data sample grows in size, it becomes possible to explore more complex networks for extracting feature representations and predicting activities. Complex architectural designs have a tendency to overfit on small-sized datasets as the model may become too closely adapted to the training set, resulting in suboptimal performance on new data instances. Overall, the dataset size and model complexity are closely tied together and should be carefully selected in consideration of each other.

6.5 Comparison with SOTA approaches

This section presents a comparison of the performance of the proposed ensemble model in relation to previously published methods for vision-based HAR. According to the findings presented in Table 15, it is evident that the ensemble model outperforms existing techniques by acquiring the best accuracy score.

In [67], Rangasamy et al. proposed a transfer-learning framework for activity recognition using the VGG-19 network pre-trained on the ImageNet dataset. The primary disadvantage of utilizing solely a pre-trained CNN for HAR is that the temporal features across frames are not taken into account, thus leading to a low accuracy score for the dataset in this study. Singh et al. [109] proposed a CRNN model for vision-based HAR. A pre-trained ResNet-152 model was combined with an RNN (or LSTM) to capture the required spatio-temporal features. Deep and Zheng [110] experimented with three state-of-the-art pre-trained CNNs – VGG-16, VGG-19, and InceptionV3. Out of these three networks,

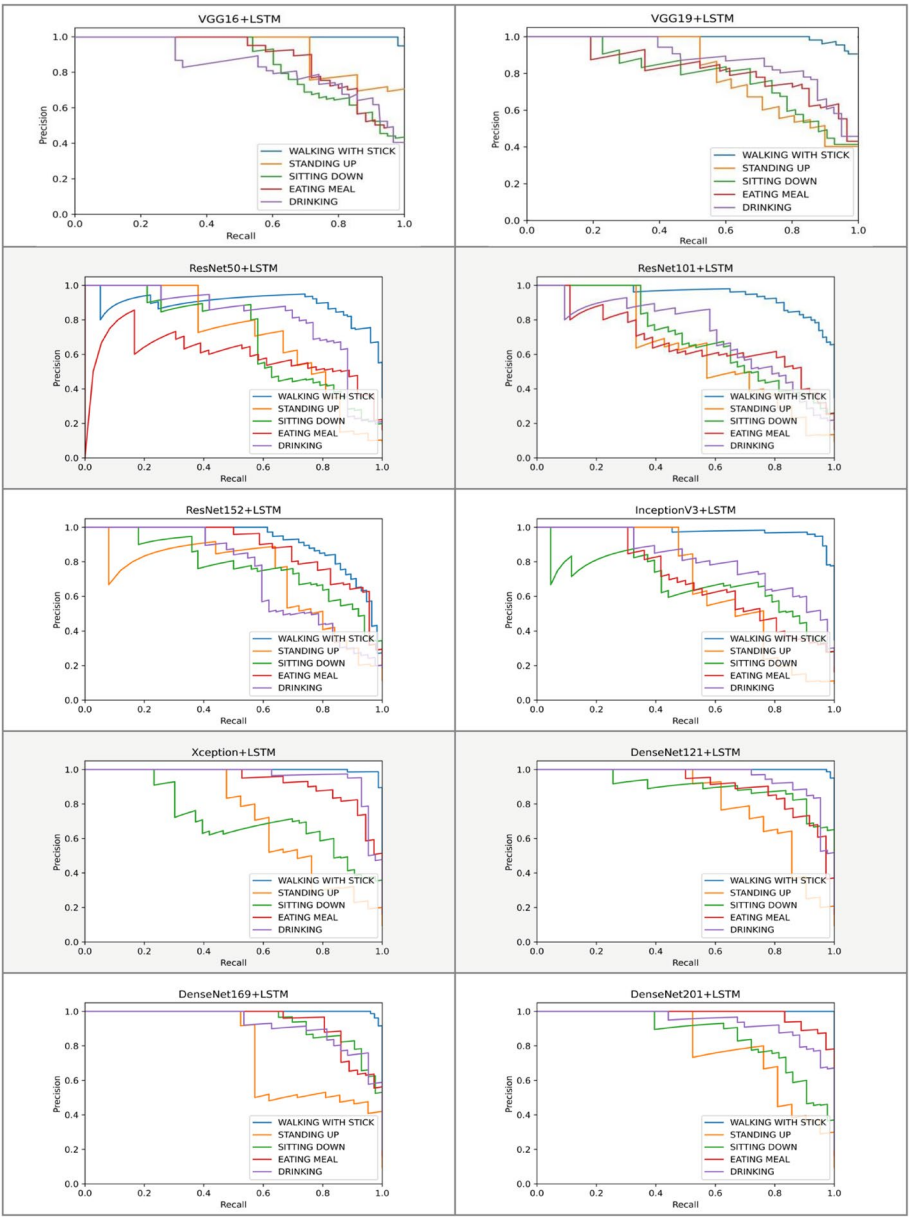


Fig. 13 Precision-recall curves for each of the network architectures

InceptionV3 helped achieve the highest accuracy score of 0.6364 on the dataset under consideration in this study.

Garg et al. [111] trained CNN and LSTM models for vision-based HAR in the KTH and Weizmann datasets. The authors concluded that the two types of models when brought together, were able to better understand and recognize activities than when they were used

Table 13 AUPRC for each of the network architectures

Network	Walking with stick	Standing up	Sitting down	Eating meal	Drinking
VGG16 + LSTM	0.9395	0.7586	0.7282	0.8437	0.5939
VGG19 + LSTM	0.9439	0.6751	0.7625	0.6712	0.7436
ResNet50 + LSTM	0.8903	0.7035	0.7115	0.5873	0.8238
ResNet101 + LSTM	0.9667	0.9038	0.9193	0.9139	0.8695
ResNet152 + LSTM	0.9203	0.4984	0.5662	0.6162	0.7052
InceptionV3 + LSTM	0.9791	0.6845	0.6881	0.7093	0.8116
Xception + LSTM	0.9975	0.7197	0.7250	0.9329	0.9624
DenseNet121 + LSTM	0.9989	0.8286	0.9009	0.9108	0.9614
DenseNet169 + LSTM	0.9986	0.8128	0.9425	0.9387	0.9253
DenseNet201 + LSTM	1.0000	0.8198	0.8589	0.9829	0.9423

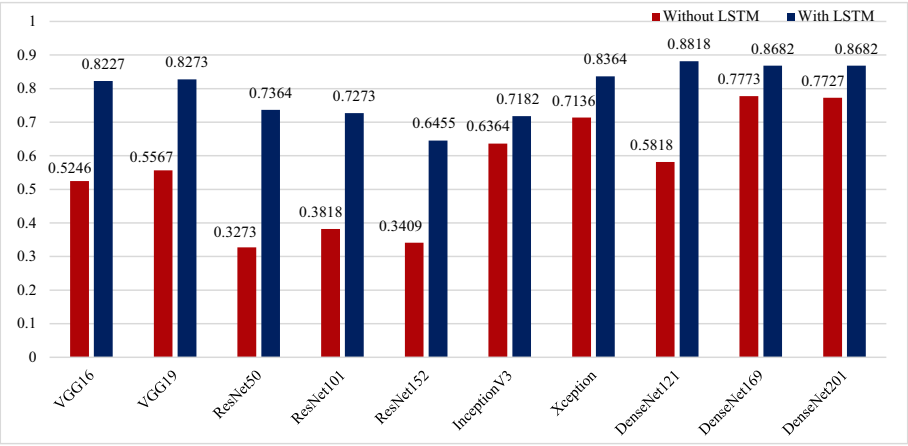


Fig. 14 Results of ablation study

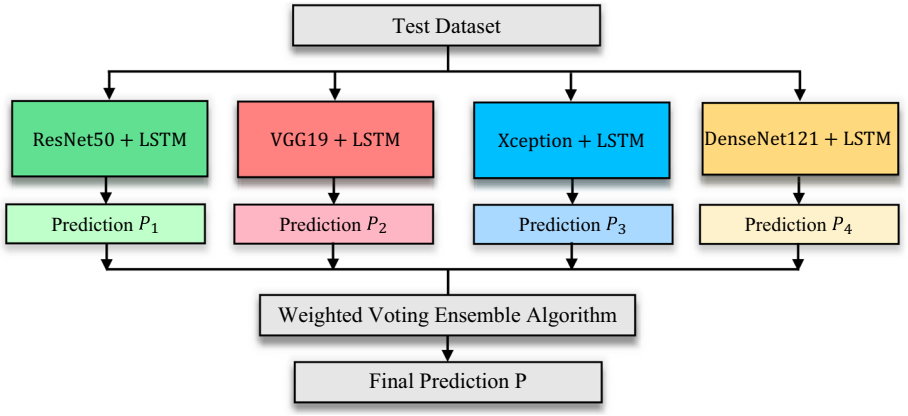


Fig. 15 Schematic map for human activity recognition via weighted voting ensemble model

Table 14 Test-set accuracy for the best weight combinations

Weights for each model				Test-set accuracy
ResNet50 + LSTM	VGG19 + LSTM	Xception + LSTM	DenseNet121 + LSTM	
0.0000	0.0000	0.0000	1.0000	0.8818
0.1000	0.1000	0.1000	0.7000	0.8909
0.2308	0.2308	0.2308	0.3076	0.9000
0.2400	0.2400	0.2400	0.2800	0.9045
0.0625	0.0625	0.4375	0.4375	0.9091
0.0588	0.0588	0.4118	0.4706	0.9136
0.2000	0.2000	0.2000	0.4000	0.9182
0.1667	0.1667	0.1667	0.4999	0.9227
0.1429	0.1429	0.2143	0.4999	0.9273
0.1333	0.1333	0.2667	0.4667	0.9318

separately. Verma and Singh [112] used VGG-19 as a feature extractor, followed by SVM as the classifier for activity recognition from the UCF Sports Action dataset. In [113], Putra et al. used the cyclical learning rate method with ResNet-50 to develop a vision-based HAR model. Kulkarni and Jadhav [114] employed VGG-16, InceptionV3, and MobileNet as feature extractors to develop a transfer-learning-based HAR model using the Collective Activity dataset [115].

6.6 Computational complexity and scalability

Computational complexity plays a critical role in evaluating the efficiency of deep learning models. The number of trainable parameters and inference time are often considered as measures of computational complexity. In this section, we analyze the trade-off between computational complexity, measured by the number of trainable parameters and inference time, and the test-set accuracy of the ten hybrid models.

Figure 16 portrays a bubble chart wherein the size of each bubble is proportional to the number of trainable parameters of the model. A larger bubble implies a greater number of trainable parameters. The test-set accuracies are plotted on the vertical axis, while the inference time (in seconds) has been specified along the horizontal axis. According to the plotted metrics, the models in the top-left region are optimal in terms of high test-set accuracy and low inference time. Despite having a lower trainable parameter count as compared to all ResNet variants, the hybrid models constituted using VGG variants achieve considerably higher test-set accuracies. The best-performing DenseNet121+LSTM model possesses an intermediate parameter count between those of models comprising VGG and ResNet variants. It is also the lowest parameter variant of the DenseNet architecture. Despite its complexity and high trainable parameter count, the ResNet architecture is unable to achieve comparable results when compared against the lower parameter alternatives such as VGG-16, VGG-19, and DenseNet-121. Upon examining the inference times, we observe that the ResNet variants, InceptionV3 and Xception, have contributed to the development of hybrid models with faster inference capability compared to those created using VGG variants. The graph clearly shows that the DenseNet 121+LSTM model, which achieves the highest test-set accuracy, also possesses a favourable balance between a low

Table 15 Comparison of proposed model with SOTA approaches

Approach	Rangasamy et al. [67]	Singh et al. [109]	Deep and Zheng [110]	Garg et al. [111]	Verma and Singh [112]	Putra et al. [113]	Kulkarni and Jadhav [114]	Proposed (Ours)
Test-set accuracy	0.5246	0.6455	0.6364	0.8455	0.6225	0.3876	0.6885	**0.9318**

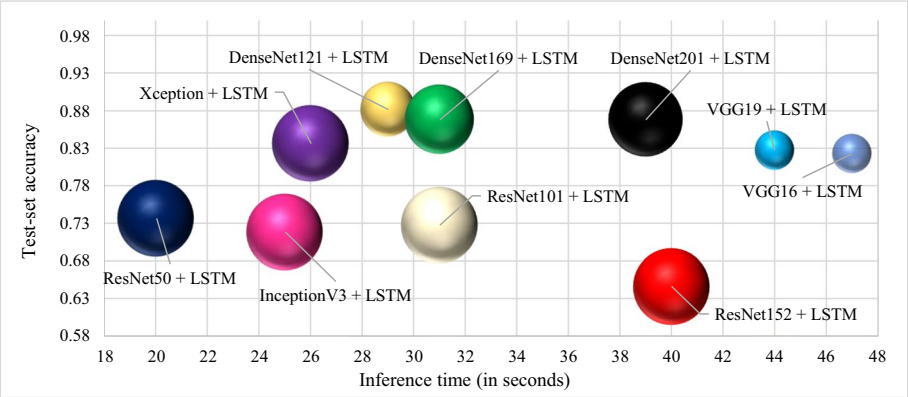


Fig. 16 Complexity analysis of the ten hybrid models analyzed in this study. The vertical axis specifies the test-set accuracy. The horizontal axis mentions the inference time (in seconds). The size of each circle is proportional to the number of trainable parameters of the model

trainable parameter count and a short inference time. Deep learning models possess a notable characteristic of scalability, enabling them to process large volumes of data effectively. Furthermore, it has been observed that larger datasets positively impact their performance. Considering this, the hybrid models and the ensemble model discussed in this paper can be applied to large-scale video datasets for human activity recognition, such as the ActivityNet and NTU datasets. Utilizing GPUs with greater processing power and larger memory capacity can significantly reduce training and inference times, thus augmenting the efficiency of the models.

7 Strengths, limitations and future research directions

The following subsections lay out the strengths of the proposed model while acknowledging its limitations that pave the way for future research endeavours.

7.1 Strengths of the proposed model

The hybrid model employs a combination of CNN and LSTM to automatically retrieve the pertinent features for human activity recognition. CNN is adept at image recognition and can extract valuable features from visual data, whereas LSTM is proficient at handling temporal information, such as the sequence of steps undertaken to perform an activity. The concurrent application of CNN and LSTM enhances the efficacy of activity recognition by accounting for both spatial and temporal connections in the visual data. Ensemble classifiers provide enhanced reliability and uniformity when applied to various subsets of the data. The stability provided by an ensemble model guarantees that the predictions are dependable and that the outcomes may be consistent, irrespective of the particular training data utilized.

7.2 Limitations, scope for improvement, and future opportunities

Though the proposed ensemble architecture helped achieve promising results for HAR, it is necessary to recognize the constraints of the current work in order to outline potential areas for future research. Firstly, the performance of the proposed method was validated using videos from a single dataset. Its performance must be tested using other publicly available video datasets for activity recognition. Secondly, the recent introduction of a variety of attention modules [116, 117] has greatly enhanced the performance and efficacy of deep learning architectures [118–120]. Future endeavours must be aimed at integrating these attention modules into the CNN-LSTM model to further facilitate activity recognition by focusing on the most crucial aspects pertaining to a particular activity. Thirdly, in order to develop generalizable architectures for HAR, it would be interesting to examine how models trained on other video datasets with overlapping activity classes perform on the STAIR Actions dataset. Lastly, keeping in mind the need to develop light-weight architectures for HAR that can easily be deployed on mobile devices, care must be taken to reduce the number of model parameters. Most current frameworks for HAR concentrate on achieving superior performance in terms of metrics like accuracy, precision, and recall but overlook the importance of model size, a critical factor for real-time deployment.

8 Conclusion

This research work experimented with several state-of-the-art pre-trained neural networks including VGG16, VGG19, ResNet50, ResNet101, ResNet152, InceptionV3, Xception, DenseNet121, DenseNet169 and DenseNet201 using transfer learning technique. The best individual model turned out to be the one trained using a combination of DenseNet121 pre-trained on the ImageNet dataset with LSTM with an accuracy of 88.18% on the test dataset, followed by DenseNet169 and DenseNet201 with an accuracy of 86.82% each. Despite the existence of numerous ensemble methods, determining the suitable ensemble configuration for a specific dataset can be challenging. The weighted voting ensemble technique is highly favoured for its straightforward concept, intuitive nature, and practical effectiveness. This technique merges a group of well-performing conceptually diverse base learners to offset their individual deficiencies and generate an improved learner. Given that the base models constituting an ensemble exhibit varying levels of performance, it is imperative to assign distinct weights to these base learners in order to enhance classification performance. On preparing a weighted voting ensemble model using the top-performing model from each category of CNN networks taken up, we obtained a test-set accuracy of 93.18%, a major improvement over the best performing individual model. This study brings out a detailed comparison and analysis and proves that human activity recognition models based on the integration of CNN with LSTM approach, when built using transfer learning methods, can give good results and save training time which would otherwise be required for training the CNN employed. The most significant aspects of our work include carrying out vision-based human activity recognition using a relatively unexplored dataset and then proposing an ensemble model after thoroughly evaluating the performance of different neural network architectures using standard metrics when incorporated as the feature extraction unit in the hybrid model described. Incorporating attention modules into the proposed

model presents promising future opportunities. With a suitable attention mechanism, the model can automatically adjust the importance of key features and reduce the influence of irrelevant noise. Light-weight architectures may be achieved, for instance, by employing computationally efficient multi-branch feature extractors such as those used in the ResNext [121] architecture.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability The STAIR Actions dataset is available at https://actions.stair.center/.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Babiker M, Khalifa OO, Htike KK, Hassan A, Zaharadeen M (2017) Automated daily human activity recognition for video surveillance using neural network. In: 2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA), pp. 1–5. https://doi.org/10.1109/ICSIMA.2017.8312024.
2. ASM, Thillaiarasu N (2022) A Survey on Different Computer Vision Based Human Activity Recognition for Surveillance Applications. In: 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1372–1376. https://doi.org/10.1109/ICCMC53470.2022.9753931
3. Kolkar R, Geetha V (2021) Human Activity Recognition in Smart Home using Deep Learning Techniques. In: 2021 13th International Conference on Information Communication Technology and System (ICTS), pp. 230–234. https://doi.org/10.1109/ICTS52701.2021.9609044
4. Niu H, Nguyen D, Yonekawa K, Kurokawa M, Wada S, Yoshihara K (2020) Multi-source Transfer Learning for Human Activity Recognition in Smart Homes. In: 2020 IEEE International Conference on Smart Computing (SMARTCOMP), pp. 274–277. https://doi.org/10.1109/SMARTCOMP50058.2020.00063
5. Siddiqi MH et al (2021) A Unified Approach for Patient Activity Recognition in Healthcare Using Depth Camera. IEEE Access 9:92300–92317. https://doi.org/10.1109/ACCESS.2021.3092403
6. Abdel-Basset M, Hawash H, Chakrabortty RK, Ryan M, Elhoseny M, Song H (2021) ST-DeepHAR: Deep Learning Model for Human Activity Recognition in IoHT Applications. IEEE Internet Things J 8(6):4969–4979. https://doi.org/10.1109/JIOT.2020.3033430
7. Vijayaprabakaran K, Sathiyamurthy K, Ponniamma M (2020) Video-Based Human Activity Recognition for Elderly Using Convolutional Neural Network. Int J Secur. Priv Pervasive Comput IJSPPC 12(1): 36–48. https://doi.org/10.4018/IJSPPC.2020010104
8. Kim K, Jalal A, Mahmood M (2019) Vision-Based Human Activity Recognition System Using Depth Silhouettes: A Smart Home System for Monitoring the Residents. J Electr Eng Technol 14:2567–2573. https://doi.org/10.1007/s42835-019-00278-8
9. D'Angelo G, Palmieri F (2021) Enhancing COVID-19 tracking apps with human activity recognition using a deep convolutional neural network and HAR-images. Neural Comput Appl.https://doi.org/10.1007/s00521-021-05913-y
10. Zhao R, Ji Q (2018) An adversarial hierarchical hidden Markov model for human pose modeling and generation. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, in AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, pp. 2636–2643. https://doi.org/10.1609/aaai.v32i1.11860
11. Lei Q, Zhang H, Xin M, Cai Y (2018) A hierarchical representation for human action recognition in realistic scenes. Multimed Tools Appl 77(9):11403–11423. https://doi.org/10.1007/s11042-018-5626-0

12. Demrozi F, Pravadelli G, Bihorac A, Rashidi P (2020) Human Activity Recognition Using Inertial, Physiological and Environmental Sensors: A Comprehensive Survey. IEEE Access 8:210816–210836. https://doi.org/10.1109/ACCESS.2020.3037715
13. Beddiar DR, Nini B, Sabokrou M, Hadid A (2020) Vision-based human activity recognition: a survey. Multimed Tools Appl 79(41):30509–30555. https://doi.org/10.1007/s11042-020-09004-3
14. Abdul Manaf F, Singh S (2021) Computer Vision-based Survey on Human Activity Recognition System, Challenges and Applications. In: 2021 3rd International Conference on Signal Processing and Communication (ICPSC), pp. 110–114. https://doi.org/10.1109/ICSPC51351.2021.9451736
15. Zhang S, Wei Z, Nie J, Huang L, Wang S, Li Z (2017) A Review on Human Activity Recognition Using Vision-Based Method. J Healthc Eng 2017:1–31. https://doi.org/10.1155/2017/3090343
16. Bhola G, Vishwakarma DK (2024) A review of vision-based indoor HAR: state-of-the-art, challenges, and future prospects. Multimed Tools Appl 83(1):1965–2005. https://doi.org/10.1007/s11042-023-15443-5
17. Jegham I, Ben Khalifa A, Alouani I, Mahjoub M (2020) Vision-based human action recognition An overview and real world challenges. Digit Investig 32: 200901 https://doi.org/10.1016/j.fsidi.2019.200901
18. Bux A (2017) Vision-based Human Action Recognition using Machine Learning Techniques 19104487 B. https://doi.org/10.17635/LANCASTER/THESIS/186
19. Ghate V, Sweetlin Hemalatha C (2023) A comprehensive comparison of machine learning approaches with hyper-parameter tuning for smartphone sensor-based human activity recognition. Meas Sens 30: 100925. https://doi.org/10.1016/j.measen.2023.100925
20. Dentamaro V, Gattulli V, Impedovo D, Manca F (2024) Human activity recognition with smartphone-integrated sensors: A survey. Expert Syst. Appl. 246:123143. https://doi.org/10.1016/j.eswa.2024.123143
21. Garcia-Gonzalez D, Rivero D, Fernandez-Blanco E, Luaces MR (2023) Deep learning models for real-life human activity recognition from smartphone sensor data. Internet Things 24:100925. https://doi.org/10.1016/j.iot.2023.100925
22. Chaurasia SK, Reddy SRN (2022) State-of-the-art survey on activity recognition and classification using smartphones and wearable sensors. Multimed Tools Appl 81(1):1077–1108. https://doi.org/10.1007/s11042-021-11410-0
23. Garain A, Dawn R, Singh S, Chowdhury C (2022) Differentially private human activity recognition for smartphone users. Multimed Tools Appl 81(28):40827–40848. https://doi.org/10.1007/s11042-022-13185-4
24. Sri Harsha NC, Anudeep YGVS, Vikash K, Ratnam DV (2021) Performance Analysis of Machine Learning Algorithms for Smartphone-Based Human Activity Recognition. Wirel Pers Commun 121(1):381–398 https://doi.org/10.1007/s11277-021-08641-7
25. Raj R, Kos A (2023) An improved human activity recognition technique based on convolutional neural network. Sci Rep 13(1):1. https://doi.org/10.1038/s41598-023-49739-1
26. Khan IU, Afzal S, Lee JW (2022) Human Activity Recognition via Hybrid Deep Learning Based Model. Sensors 22(1):323. https://doi.org/10.3390/s22010323
27. Xia K, Huang J, Wang H (2020) LSTM-CNN Architecture for Human Activity Recognition. IEEE Access 8:56855–56866. https://doi.org/10.1109/ACCESS.2020.2982225
28 Guan Y, Plötz T (2017) Ensembles of Deep LSTM Learners for Activity Recognition using. Wearables Proc ACM Interact Mob Wearable Ubiquitous Technol 1(2):11:1-11:28. https://doi.org/10.1145/3090076
29. Pienaar SW, Malekian R (2019) Human Activity Recognition using LSTM-RNN Deep Neural Network Architecture. In: 2019 IEEE 2nd Wireless Africa Conference (WAC), pp. 1–5. https://doi.org/10.1109/AFRICA.2019.8843403
30 Ordóñez FJ, Roggen D (2016) Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors 16(1):1. https://doi.org/10.3390/s16010115
31. Kaya Y, Topuz EK (2024) Human activity recognition from multiple sensors data using deep CNNs. Multimed Tools Appl 83(4):10815–10838. https://doi.org/10.1007/s11042-023-15830-y
32. Park J, Lim W-S, Kim D-W, Lee J (2023) GTSNet: Flexible architecture under budget constraint for real-time human activity recognition from wearable sensor. Eng Appl Artif Intell 124:106543. https://doi.org/10.1016/j.engappai.2023.106543
33. Matey-Sanz M, Casteleyn S, Granell C (2023) Dataset of inertial measurements of smartphones and smartwatches for human activity recognition. Data Brief 51:109809. https://doi.org/10.1016/j.dib.2023.109809

34. Zhang L, Yu J, Gao Z, Ni Q (2024) A multi-channel hybrid deep learning framework for multi-sensor fusion enabled human activity recognition. Alex Eng J 91:472–485. https://doi.org/10.1016/j.aej.2024.01.030
35. Ullah S, Pirahandeh M, Kim D-H (2024) Self-attention deep ConvLSTM with sparse-learned channel dependencies for wearable sensor-based human activity recognition. Neurocomputing 571:127157. https://doi.org/10.1016/j.neucom.2023.127157
36. Lalwani P, Ramasamy G (2024) Human activity recognition using a multi-branched CNN-BiLSTM-BiGRU model. Appl Soft Comput 154:111344. https://doi.org/10.1016/j.asoc.2024.111344
37. Sansano E, Montoliu R, Belmonte Fernández Ó (2020) A study of deep neural networks for human activity recognition. Comput Intell 36(3): 1113–1139. https://doi.org/10.1111/coin.12318
38. Mallik M, Sarkar G, Chowdhury C (2023) A Deep Learning Framework for Smartphone Based Human Activity Recognition. Mob Netw Appl. https://doi.org/10.1007/s11036-023-02117-7
39. Serpush F, Menhaj MB, Masoumi B, Karasfi B (2023) Wearable Sensors-Based Human Activity Recognition with Deep Convolutional Neural Network and Fuzzy Classification. Wirel Pers Commun 133(2):889–911. https://doi.org/10.1007/s11277-023-10797-3
40. Noori FM, Riegler M, Uddin MZ, Torresen J (2020) Human Activity Recognition from Multiple Sensors Data Using Multi-fusion Representations and CNNs. ACM Trans Multimed Comput Commun Appl 16 2 45 1-45. https://doi.org/10.1145/3377882
41. Jameer S, Syed H (2023) A DCNN-LSTM based human activity recognition by mobile and wearable sensor networks. Alex Eng J 80:542–552. https://doi.org/10.1016/j.aej.2023.09.013
42. Bagci Das D and Birant D (2023) Human activity recognition based on multi-instance learning. Expert Syst 40(7):e13256. https://doi.org/10.1111/exsy.13256
43. Kumar LMA, Murugan S (2023) Design of cuckoo search optimization with deep belief network for human activity recognition and classification. Multimed Tools Appl 82(19):29823–29841. https://doi.org/10.1007/s11042-023-14977-y
44. Choudhury NA, Soni B (2023) In-depth analysis of design & development for sensor-based human activity recognition system. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-16423-5
45. Imanzadeh S, Tanha J, Jalili M (2024) Ensemble of deep learning techniques to human activity recognition using smart phone signals. Multimed Tools Appl. https://doi.org/10.1007/s11042-024-18935-0
46. Thakur U, Prajapati A, Vidyarthi A (2024) A bilateral assessment of human activity recognition using grid search based nonlinear multi-task least squares twin support vector machine. Multimed Tools Appl. https://doi.org/10.1007/s11042-024-18993-4
47. Boudjema A, Titouna F, Titouna C (2023) AReNet: Cascade learning of multibranch convolutional neural networks for human activity recognition. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-17496-y
48. Gholamrezaii M, AlModarresi S (2021) A time-efficient convolutional neural network model in human activity recognition. Multimed Tools Appl 80(13):19361–19376. https://doi.org/10.1007/s11042-020-10435-1
49. Geravesh S, Rupapara V (2023) Artificial neural networks for human activity recognition using sensor based dataset. Multimed Tools Appl 82(10):14815–14835. https://doi.org/10.1007/s11042-022-13716-z
50. Hasan MA, Li F, Piet A, Gouverneur P, Irshad MT, Grzegorzek M (2023) Exploring the Benefits of Time Series Data Augmentation for Wearable Human Activity Recognition. In: Proceedings of the 8th international Workshop on Sensor-Based Activity Recognition and Artificial Intelligence, in iWOAR '23. New York, NY, USA: Association for Computing Machinery, pp. 1–7. https://doi.org/10.1145/3615834.3615842
51. Raja Sekaran S, Han PY, Yin OS (2023) Smartphone-based human activity recognition using lightweight multiheaded temporal convolutional network. Expert Syst Appl 227:120132. https://doi.org/10.1016/j.eswa.2023.120132
52. Thakur D, Biswas S (2024) Permutation importance based modified guided regularized random forest in human activity recognition with smartphone. Eng Appl Artif Intell 129:107681. https://doi.org/10.1016/j.engappai.2023.107681
53. Ghalan M, Aggarwal RK (2022) Daily life human activities recognition using a novel AR-DenseNet. Concurr Comput Pract Exp 34(22):e7142. https://doi.org/10.1002/cpe.7142
54. Ouyang K, Pan Z (2023) Multi-model weighted voting method based on convolutional neural network for human activity recognition. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-17500-5
55. Zhang Z, Wang W, An A, Qin Y, Yang F (2023) A human activity recognition method using wearable sensors based on convtransformer model. Evol Syst 14(6):939–955. https://doi.org/10.1007/s12530-022-09480-y

56. Tanigaki K, Teoh TC, Yoshimura N, Maekawa T, Hara T (2022) Predicting Performance Improvement of Human Activity Recognition Model by Additional Data Collection. Proc ACM Interact Mob Wearable Ubiquitous Technol 6(3): 142:1-142:33, https://doi.org/10.1145/3550319
57. Dua N, Singh SN, Semwal VB, Challa SK (2023) Inception inspired CNN-GRU hybrid network for human activity recognition. Multimed Tools Appl 82(4):5369–5403. https://doi.org/10.1007/s11042-021-11885-x
58. Ankalaki S, Thippeswamy MN (2023) A novel optimized parametric hyperbolic tangent swish activation function for 1D-CNN: application of sensor-based human activity recognition and anomaly. Multimed. detection'https://doi.org/10.1007/s11042-023-15766-3
59. Prabono AG, Yahya BN, Lee S-L (2022) Multiple-instance domain adaptation for cost-effective sensor-based human activity recognition. Future Gener Comput Syst 133:114–123. https://doi.org/10.1016/j.future.2022.03.006
60. Mhalla A, Favreau J-M (2024) Domain adaptation framework for personalized human activity recognition models. Multimed Tools Appl. https://doi.org/10.1007/s11042-024-18267-z
61. Zebhi S (2024) An efficient 3D convolutional neural network with informative 3D volumes for human activity recognition using wearable sensors. Multimed Tools Appl 83(14):42233–42256. https://doi.org/10.1007/s11042-023-17400-8
62. Lattanzi E, Calisti L, Capellacci P (2023) Lightweight accurate trigger to reduce power consumption in sensor-based continuous human activity recognition. Pervasive Mob Comput 96:101848. https://doi.org/10.1016/j.pmcj.2023.101848
63. Liang J, Zhang L, Bu C, Yang G, Wu H, Song A (2024) Plug-and-play multi-dimensional attention module for accurate Human Activity Recognition. Comput Net 244:110338. https://doi.org/10.1016/j.comnet.2024.110338
64. Pareek G, Nigam S, Singh R (2024) Modeling transformer architecture with attention layer for human activity recognition. Neural Comput Appl 36(10):5515–5528. https://doi.org/10.1007/s00521-023-09362-7
65. Bodhe R, Sivakumar S, Sakarkar G, Juwono FH, Apriono C (2024) Outdoor activity classification using smartphone based inertial sensor measurements. Multimed Tools Appl. https://doi.org/10.1007/s11042-024-18599-w
66. Stuchbury-Wass J, Ferlini A, Mascolo C (2023) Multimodal Attention Networks for Human Activity Recognition From Earable Devices. In Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers, in UbiComp/ISWC '22 Adjunct. New York, NY, USA: Association for Computing Machinery. pp. 258–260. https://doi.org/10.1145/3544793.3563422
67. Rangasamy K, As'ari MA, Rahmad NA, Ghazali NF (2020) Hockey activity recognition using pre-trained deep learning model. ICT Express 6(3):170–174. https://doi.org/10.1016/j.icte.2020.04.013
68. Hutchison D et al. (2010) Action Classification in Soccer Videos with Long Short-Term Memory Recurrent Neural Networks. In: Artificial Neural Networks – ICANN 2010, vol. 6353, K. Diamantaras, W. Duch, and L. S. Iliadis, (Eds), In: Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 154–159. https://doi.org/10.1007/978-3-642-15822-3_20
69. Grushin A, Monner DD, Reggia JA, Mishra A (2013) Robust human action recognition via long short-term memory. In: The 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA: IEEE, pp. 1–8. https://doi.org/10.1109/IJCNN.2013.6706797
70. Veeriah V, Zhuang N, Qi G-J (2015) Differential Recurrent Neural Networks for Action Recognition. In: presented at the 2015 IEEE International Conference on Computer Vision (ICCV), IEEE Computer Society, pp. 4041–4049. https://doi.org/10.1109/ICCV.2015.460.
71. Donahue J et al (2017) Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. IEEE Trans Pattern Anal Mach Intell 39(4):677–691. https://doi.org/10.1109/TPAMI.2016.2599174
72. Carreira J, Zisserman A (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733. https://doi.org/10.1109/CVPR.2017.502
73. Ng JY-H, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: Deep networks for video classification. In: Presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, pp. 4694–4702. https://doi.org/10.1109/CVPR.2015.7299101
74. Qiu Z, Yao T, Mei T (2017) Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5534–5542. https://doi.org/10.1109/ICCV.2017.590

75. Núñez JC, Cabido R, Pantrigo JJ, Montemayor AS, Vélez JF (2018) Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition. Pattern Recognit 76:80–94. https://doi.org/10.1016/j.patcog.2017.10.033
76. Pandya M, Pillai A, Rupani H (2020) Segregating and Recognizing Human Actions from Video Footages Using LRCN Technique, pp. 3–13. https://doi.org/10.1007/978-981-15-3383-9_1
77. K. Soomro, A. R. Zamir, and M. Shah (2012) UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild'. arXiv. Accessed: 17 May 2022. http://arxiv.org/abs/1212.0402
78. Negi A, Kumar K (2023) End-to-end residual learning-based deep neural network model deployment for human activity recognition. Int J Multimed Inf Retr 12(1):1. https://doi.org/10.1007/s13735-023-00269-6
79. Nguyen HP, Ribeiro B (2023) Video action recognition collaborative learning with dynamics via PSO-ConvNet. Transformer Sci Rep 13(1):1. https://doi.org/10.1038/s41598-023-39744-9
80. Ge H, Yan Z, Yu W, Sun L (2019) An attention mechanism based convolutional LSTM network for video action recognition. Multimed Tools Appl 78(14):20533–20556. https://doi.org/10.1007/s11042-019-7404-z
81. Vishwakarma DK, Dhiman C (2019) A unified model for human activity recognition using spatial distribution of gradients and difference of Gaussian kernel. Vis Comput 35(11):1595–1613. https://doi.org/10.1007/s00371-018-1560-4
82. Nasir IM, Raza M, Shah JH, Wang S-H, Tariq U, Khan MA (2022) HAREDNet: A deep learning based architecture for autonomous video surveillance by recognizing human actions. Comput Electr Eng 99:107805. https://doi.org/10.1016/j.compeleceng.2022.107805
83. Dharmalingam S, Palanisamy A (2018) Vector space based augmented structural kinematic feature descriptor for human activity recognition in videos. ETRI J 40(4):499–510. https://doi.org/10.4218/etrij.2018-0102
84. Shrivastava R, Tiwari V, Jain S, Tiwari B, Kushwaha AKS, Singh VP (2022) A role-entity based human activity recognition using inter-body features and temporal sequence memory. IET Image Process 16(11):2911–2921. https://doi.org/10.1049/ipr2.12472
85. Kushwaha A, Srivastava P, Khare A (2023) Human activity recognition based on integration of multilayer information of convolutional neural network architecture. Concurr Comput Pract Exp 35(5):e7571. https://doi.org/10.1002/cpe.7571
86. Zhao Y et al (2023) Infusing a Convolutional Neural Network with Encoded Joint Node Image Data to Recognize 25 Daily Human Activities. Adv Intell Syst 5(11):2300266. https://doi.org/10.1002/aisy.202300266
87. Kushwaha A, Khare A, Prakash O (2024) Human activity recognition algorithm in video sequences based on the fusion of multiple features for realistic and multi-view environment. Multimed Tools Appl 83(8):22727–22748. https://doi.org/10.1007/s11042-023-16364-z
88. Vishwakarma DK, Jain K (2022) Three-dimensional human activity recognition by forming a movement polygon using posture skeletal data from depth sensor. ETRI J 44(2):286–299. https://doi.org/10.4218/etrij.2020-0101
89. Riahi M, Eslami M, Safavi SH, Torkamani Azar F (2020) Human activity recognition using improved dynamic image. IET Image Process 14(13):3223–3231. https://doi.org/10.1049/iet-ipr.2019.1739
90. Samma H, Sama ASB (2024) Optimized deep learning vision system for human action recognition from drone images. Multimed Tools Appl. 83(1):1143–1164. https://doi.org/10.1007/s11042-023-15930-9
91. Alfasly S et al (2023) FastPicker: Adaptive independent two-stage video-to-video summarization for efficient action recognition. Neurocomputing 516:231–244. https://doi.org/10.1016/j.neucom.2022.10.037
92. Varshney N, Bakariya B (2022) Deep convolutional neural model for human activities recognition in a sequence of video by combining multiple CNN streams. Multimed Tools Appl 81(29):42117–42129. https://doi.org/10.1007/s11042-021-11220-4
93. Sun X, Wang Y, Khan J (2023) Hybrid LSTM and GAN model for action recognition and prediction of lawn tennis sport activities. Soft Comput 27(23):18093–18112. https://doi.org/10.1007/s00500-023-09215-4
94. Sinha KP, Kumar P (2023) Human activity recognition from uav videos using an optimized hybrid deep learning model. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-17289-3
95. Kushwaha A, Khare A, Prakash O (2023) Micro-network-based deep convolutional neural network for human activity recognition from realistic and multi-view visual data. Neural Comput Appl 35(18):13321–13341. https://doi.org/10.1007/s00521-023-08440-0

96. Singh R, Khurana R, Kushwaha AKS, Srivastava R (2020) Combining CNN streams of dynamic image and depth data for action recognition. Multimed Syst 26(3):313–322. https://doi.org/10.1007/s00530-019-00645-5
97. Gupta N, Gupta SK, Pathak RK, Jain V, Rashidi P, Suri JS (2022) Human activity recognition in artificial intelligence framework: a narrative review. Artif Intell Rev 55(6):4755–4808. https://doi.org/10.1007/s10462-021-10116-x
98. Aggarwal S, Pandey A, Vishwakarma DK (2023) Multimodal Sarcasm Recognition by Fusing Textual, Visual and Acoustic content via Multi-Headed Attention for Video Dataset. In: 2023 World Conference on Communication & Computing (WCONF), pp. 1–5. https://doi.org/10.1109/WCONF58270.2023.10235179.
99. Aggarwal S, Vishwakarma DK (2023) Protecting our Children from the Dark Corners of YouTube: A Cutting-Edge Analysis. In: 2023 4th IEEE Global Conference for Advancement in Technology (GCAT), pp. 1–5. https://doi.org/10.1109/GCAT59970.2023.10353306
100. Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
101. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.
102. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun (eds). Accessed: 03 Mar 2024. http://arxiv.org/abs/1409.1556
103. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
104. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the Inception Architecture for Computer Vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826. https://doi.org/10.1109/CVPR.2016.308
105. Chollet F (2017) Xception: Deep Learning with Depthwise Separable Convolutions. In: presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, pp. 1800–1807. https://doi.org/10.1109/CVPR.2017.195
106. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243
107. Yoshikawa Y, Lin J, Takeuchi A (2018) STAIR Actions: A Video Dataset of Everyday Home Actions, ArXiv180404326 Cs. Accessed: 12 May 2022. http://arxiv.org/abs/1804.04326
108. Zhang Z, Sabuncu M (2018) Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In: Advances in Neural Information Processing Systems, Curran Associates, Inc. Accessed: 03 Mar 2024. https://proceedings.neurips.cc/paper_files/paper/2018/hash/f2925f97bc13ad2852a7a551802feea0-Abstract.html
109. Singh R, Kushwaha AKS, Khurana R, Srivastava R. (2019) Activity Recognition by Delving deeper using CNN and RNN. In: 2019 4th International Conference on Information Systems and Computer Networks (ISCON), pp. 610–614. https://doi.org/10.1109/ISCON47742.2019.9036262
110. Deep S, Zheng X (2019) Leveraging CNN and Transfer Learning for Vision-based Human Activity Recognition. In: 2019 29th International Telecommunication Networks and Applications Conference (ITNAC), pp. 1–4. https://doi.org/10.1109/ITNAC46935.2019.9078016
111. Garg A, Nigam S, Singh R (2022) Vision based Human Activity Recognition using Hybrid Deep Learning. In: 2022 International Conference on Connected Systems & Intelligence (CSI), pp. 1–6. https://doi.org/10.1109/CSI54720.2022.9924016
112. Verma KK, Mohan Singh B (2021) Vision based Human Activity Recognition using Deep Transfer Learning and Support Vector Machine. In: 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), pp. 1–9. https://doi.org/10.1109/UPCON52273.2021.9667661
113. Putra RK, Rachmawati E, Sthevanie F (2021) Vision-Based Employee Activity Classification. In: 2021 9th International Conference on Information and Communication Technology (ICoICT), pp. 37–41. https://doi.org/10.1109/ICoICT52021.2021.9527492
114. Kulkarni SS, Jadhav S (2023) Insight on Human Activity Recognition Using the Deep Learning Approach. In: 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 1–5. https://doi.org/10.1109/ESCI56872.2023.10099759

115. 'Collective Activity Dataset'. Accessed: 09 Mar 2024. https://cvgl.stanford.edu/projects/collective/collectiveActivity.html
116. Hou Q, Zhou D, Feng J (2021) Coordinate Attention for Efficient Mobile Network Design. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13708–13717. https://doi.org/10.1109/CVPR46437.2021.01350.
117. Wang Q, Wu B, Zhu P, Li P, Zuo W, Hu Q (2020) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In: Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, pp. 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155.
118. Chhabra A, Vishwakarma DK (2023) Multimodal hate speech detection via multi-scale visual kernels and knowledge distillation architecture. Eng Appl Artif Intell 126:106991. https://doi.org/10.1016/j.engappai.2023.106991
119. Pandey A, Vishwakarma DK (2023) VABDC-Net: A framework for Visual-Caption Sentiment Recognition via spatio-depth visual attention and bi-directional caption processing. Knowl-Based Syst 269: 110515. https://doi.org/10.1016/j.knosys.2023.110515
120. Yadav A, Vishwakarma DK (2024) AW-MSA: Adaptively weighted multi-scale attentional features for DeepFake detection. Eng Appl Artif Intell 127:107443. https://doi.org/10.1016/j.engappai.2023.107443
121. Xie S, Girshick R, Dollar P, Tu Z, He K (2017) Aggregated Residual Transformations for Deep Neural Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI: IEEE, pp. 5987–5995. https://doi.org/10.1109/CVPR.2017.634

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Cluster Computing
https://doi.org/10.1007/s10586-024-04550-z

Workload dynamics implications in permissioned blockchain scalability and performance

Anita Thakur[1] · Virender Ranga[1] · Ritu Agarwal[1]

Received: 20 December 2023 / Revised: 30 April 2024 / Accepted: 7 May 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
Blockchain Technology has grown exponentially in recent years due to its decentralized, immutable, transparent data storage, transaction sharing, and processing capabilities. With the emergence of different blockchain platforms, it is important to analyze and evaluate the performance of these platforms in various scenarios. The popularity of the public blockchain, i.e., Bitcoin and Ethereum, has increased manifold. But in distinction to a public blockchain, there is a permissioned and private blockchain that allows restricted involvement of users in the network. To make a well-informed decision regarding the selection of an appropriate platform for utilization, it is crucial to evaluate diverse performance metrics among the numerous available blockchain platforms. In this study, we assessed the performance of the Hyperledger Fabric blockchain (HLF), taking into account various metrics such as resource consumption, throughput, success rate, and latency. We have incorporated parameters such as the ordering service, programming language to write chaincode/smart contracts, number of transactions, transactions per second, and organizations to evaluate the system's performance. Along with our analysis, we also suggested potential areas of research for future development of blockchain technology.

Keywords Hyperledger fabric · Ordering services · Performance metrics · Hyperledger caliper · Chaincode · Benchmarking

1 Introduction

One of the most ground-breaking and ingenious advancements in recent years is blockchain technology. Blockchain is a solid platform that has a wide range of potential uses across many industries. It was first developed as the underpinning technology for the cryptocurrency Bitcoin. With no need for a central authority or middleman, blockchain technology essentially functions as a decentralized, distributed ledger that is well-kept and verified by a network of users. Increased security, transparency, efficiency, and cost-effectiveness are just a few advantages that this technology offers. Blockchain is a topic of tremendous significance for scholars, developers, and business leaders due to its potential to positively change industries, including finance, healthcare, and more. The launch of the Bitcoin network in 2009 helped popularise the technology that has subsequently been employed in many other cryptocurrencies [1].

Through a distributed system, these digital currencies enable the movement of electronic money. Users can use digital signatures to give others ownership of their data, and the transactions are publicly visible on the Bitcoin blockchain. This makes it possible for any user on the network to confirm the validity of transactions independently. The blockchain's decentralized structure makes it possible for the system to work without having a single point of failure that could cause the database as a whole to crash. In a blockchain system, a copy of the ledger is present on every computer in the network at once. This differs from a traditional business where client information is stored on a server farm in a single building. Data loss is a risk even with server farms if proper backups aren't

✉ Anita Thakur
anitathakur_2k21phdit504@dtu.ac.in

Virender Ranga
drvirender.ranga@gmail.com

Ritu Agarwal
ritu.jeea@gmail.com

[1] Department of Information Technology, Delhi Technological University, Delhi 110042, India

Published online: 31 May 2024

available. As a result, the blockchain can continue to function even if one or more nodes go offline. When new data, such as new transactions, are added to the blockchain network, they are first broadcast to the nodes on the network. The nodes then validate the transactions and add them to a pool of unconfirmed transactions. The miners then choose transactions from this pool and add them to the next block they are mining. The blockchain gets its name from the way these blocks are connected to create a chain of documents. When data is stored on a blockchain, it becomes immutable, meaning it cannot be altered or modified once added. To incorporate new information, additional blocks must be appended to the existing chain. Beyond just supporting Bitcoin, blockchain technology's fundamental objective is to deliver third-party advantages like non-repudiation, security, authenticity, and integrity in a decentralized and distributed setting [2].

It is appropriate for any transactional system utilized by numerous independent, trustless parties since it provides auditability, transparency, immutability (transactions cannot be altered once published), and pseudonymity to users. These characteristics apply to any systems that need a high level of decentralized trust, security, and dependability in addition to monetary systems. By imposing restrictions on their adoption, permissioned blockchains enable users to alter the behavior of network nodes. The number of users who can validate transactions is restricted on permissioned blockchains, in contrast to conventional blockchains, which can be completely open and decentralized [3].

Furthermore, alterations in protocols that enable the execution of Turing-complete codes have simplified the creation of distributed applications that are based on smart contracts. Yet, fully permissioned blockchains are similar to conventional shared databases in many aspects. Thus, it's possible that this intricate design is not required in all circumstances. Although blockchain technology has a lot of capabilities for the development of future internet systems, there are still several technical issues that need to be resolved. Due to their mining-based consensus process, poor transaction throughput, and delaying transaction finality, public blockchains like Bitcoin and other POW consensus-based blockchains experience performance and scalability concerns [4] [5]. In September 2022, the Merge represented the Ethereum network's transition from a mining-based consensus mechanism, known as Proof-of-Work (PoW), to a Proof-of-Stake (PoS) consensus mechanism. So, after the merge, Ethereum no longer relies on mining to achieve consensus. Ethereum, a prominent decentralized platform, has addressed scalability challenges through a multifaceted approach. While Layer 2 scaling solutions, such as rollups, currently play a significant role, future advancements like Danksharding and proto-danksharding promise even greater transaction throughput and cost reductions. Private permissioned blockchain platforms are utilized for Business-to-Business (B2B) and Business-to-Consumer (B2C) transactions between known parties as they provide higher throughput and more effective consensus techniques. In B2B interactions, businesses often share sensitive data like pricing models or trade secrets. Private blockchains provide a permissioned environment where only authorized participants can access this data, reducing the risk of breaches. While B2C applications are still emerging, private blockchains can offer value in specific areas like loyalty programs and product verification. To decrease transaction processing delay and boost performance and scalability, the DLT-based blockchain 3.0 platforms use cutting-edge validation and voting procedures in addition to data structures like Directed Acyclic Graphs (DAG).

One of the most critical issues is scalability, particularly concerning the limiting block size of the Bitcoin network, which is currently fixed at 1 MB for each block. Because of this, the transaction rate is limited to only 7 Transactions Per Second (TPS), which makes it unfit for high-frequency trading. As a small number of users are willing to sustain such a huge blockchain, increasing the block size can result in centralization by slowing network propagation and increasing storage needs. Hence, it is necessary to address the trade-off between block size and security. Another major issue is the potential for selfish mining tactics, such as hiding mined blocks for future gain, to let miners earn more money than is reasonable [6]. This could lead to a lot of branch occurrences, which would be bad for blockchain growth. Therefore, solutions are required to deal with this problem. Even when users exclusively conduct transactions using their public and private keys, there is a chance of privacy leakage.

Blockchain's need for anonymity introduces a trust risk because dishonest users may take advantage of it. To safeguard the integrity of the ledger, each transaction needs to be confirmed before being added to a block. Consensus algorithms are used to decide which blocks to add to the blockchain, taking advantage of users' common desire to keep the system honest. As the backbone of all blockchain transactions, consensus algorithms are crucial in building trust and ensuring the correct storage of transaction data [7]. In a blockchain network, a consensus mechanism is a collection of guidelines and procedures that enables nodes to concur on the ledger's current state. Each node in a blockchain network maintains a copy/duplicate of the ledger, which is dispersed throughout the network. Each node has the power to validate and add new transactions to the ledger. There is no central body in a decentralized network to validate or confirm the authenticity of transactions. This is where consensus mechanisms come into play, as they provide a way for nodes to reach an agreement on

the state of the ledger without the need for a centralized authority.

Consensus algorithms are an important aspect of blockchain technology because they enable network nodes to agree on the validity of transactions and the order in which they are added to the blockchain. These algorithms act as a safeguard to ensure the dependability and validity of transactions. Proof of Stake (PoS), Delegated Proof of Stake (DPoS) [8], Proof of Work (PoW) [9], and Proof of Activity (PoA) are the most commonly used consensus algorithms in public blockchain applications. However, it is important to note that each of these algorithms has distinct characteristics in computing cost, security, and consensus efficiency. Unlike public blockchains, where anyone can participate in the consensus process, permissioned blockchains only restrict participation to authorized parties. This allows for more control over who validates transactions on the network. Permissioned blockchains typically outperform public blockchains due to restricted participation. The performance of a permissioned blockchain network can be quantified in terms of transaction throughput, confirmation latency, and scalability.

1.1 Motivation

Many ways to improve the data protection and verbosity of contemporary information systems are provided by blockchain technology. Regarding customization and potential use cases, the HLF offers a variety of functions. It permits a developer to create a unique business network on blockchain that comprises various heterogeneous clients, each having a certain purpose. The motivation behind conducting this study is to identify the various areas where blockchain platforms can be made more efficient depending on the performance measure. With the analysis of the performance, it can be easy to identify the bottleneck and the factors affecting the performance of the blockchain platform.

1.2 Problem statement

Although blockchain technology presents an enticing substitute for centralized database systems with its emphasis on- security and transparency, its practicality is frequently hindered by performance limitations. Issues such as inefficiencies in transaction processing and the absence of standardized protocols can serve as substantial barriers to the widespread adoption of blockchain. These constraints materialize as scalability hurdles, increased transaction latency, and bottlenecks in throughput. Blockchains with limited scalability struggle to handle the high volume of transactions typical in enterprise environments. For example, Bitcoin processes around 7 TPS due to its small block size and Proof-of-Work consensus mechanism. This is insufficient for large-scale adoption as a digital payment system. While faster than Bitcoin, Ethereum still experiences congestion during periods of high activity, leading to an increase in latency. Also, previous works have primarily focused on delineating the architectural and conceptual frameworks for implementing blockchain across different domains. Nonetheless, the successful deployment of blockchain requires careful consideration of factors such as the number of transactions, transaction rates, and the number of participants. These variables can significantly influence performance metrics such as success rate, throughput, resource consumption, latency, and the function relating to the creation and querying of transactions.

1.3 Research contributions

Here, we have presented the key contribution of our work:

1. The work concentrates on evaluating the performance of Hyperledger Fabric, analyzing multiple key performance indicators to gauge the effectiveness of the blockchain platform. This research builds upon our prior work [10], which delved into the performance of three prominent blockchain platforms: Hyperledger Fabric, privately deployed Ethereum, and Hyperledger Besu. A comparison of the platforms revealed that Hyperledger Fabric surpasses Ethereum and Hyperledger Besu in terms of throughput and latency. This paper presents a comprehensive analysis specifically focusing on the performance attributes of Hyperledger Fabric.
2. The work examined the impact of increased workload, TPS, ordering services, and organizational factors on the scalability and performance of the blockchain platform.
3. The work investigates the various errors that lead to transaction failure.

The outline of the study is illustrated (Fig. 1) as follows. A general introduction to blockchain, its application, and its drawbacks are described in section one.

The survey of existing literature is provided in section two. Section three describes the tools and techniques to measure the performance and scalability of the HLF blockchain. Section four outlines the experimental design and reports the findings. Section five discusses the experimental outcome and our analysis. Section six of the paper summarizes the findings and conclusions drawn from the research and outlines potential areas for future work.

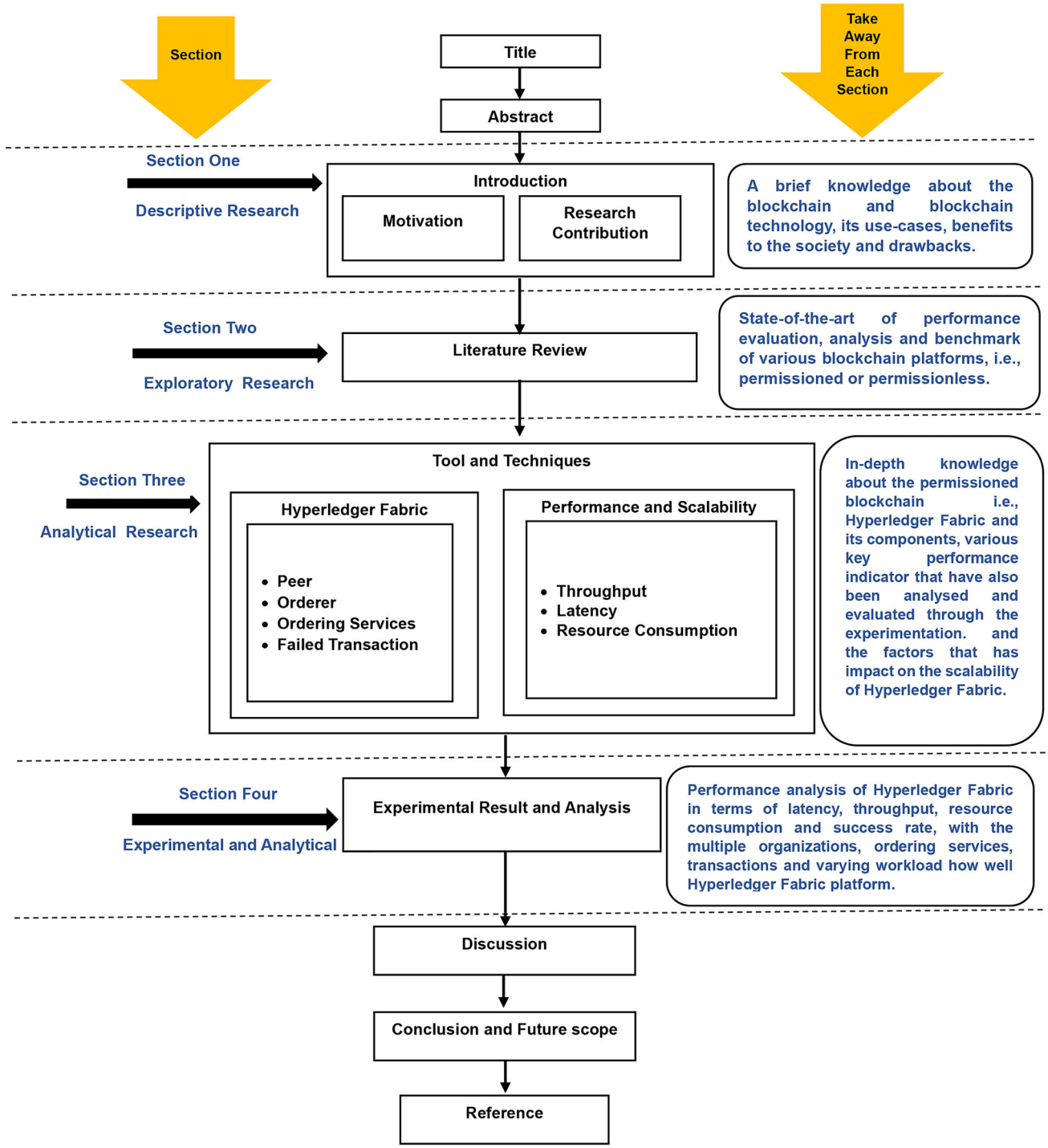


Fig. 1 Structure of our study

2 Literature review

This research employed a literature review to identify relevant scholarly work. We conducted an exhaustive search of print sources and publications across diverse publishers, encompassing works published between 2017 and 2023. To ensure the quality and relevance of the included studies, we used a rigorous methodological framework utilizing pre-defined exclusion and inclusion criteria, as illustrated in Fig. 2. Many performance evaluations have been carried out on the blockchain platform to gauge its effectiveness and pinpoint possible areas for improvement. These studies often employ benchmarking software to assess the platform's throughput and latency

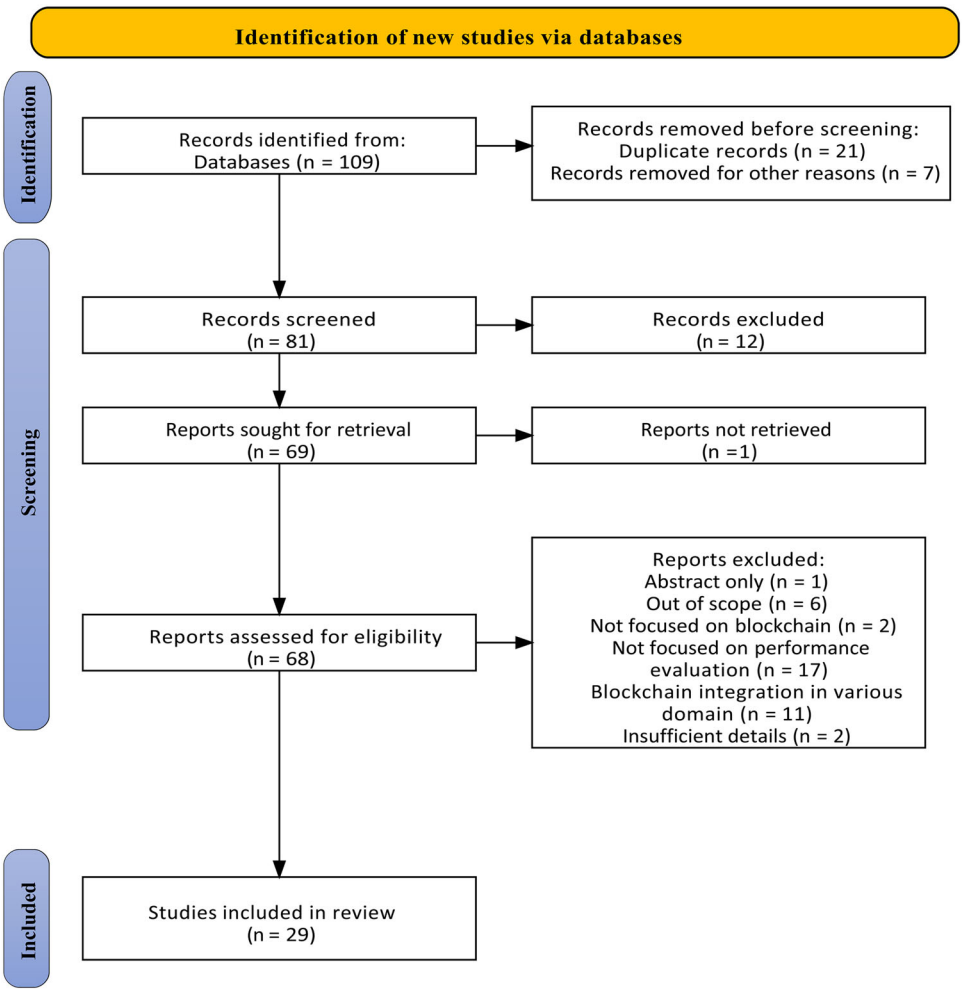


Fig. 2 PRISMA flow diagram for selection of studies

under several scenarios and identify the variables influencing its performance. Here, Table 1 outlines relevant studies concerning the performance analysis of blockchain platforms alongside our research, which encompasses various parameters.

2.1 Studies related to hyperledger fabric

To analyze the performance of HLF version 1.0 and version 0.6 and compare them in the article using a variety of measures, including latency, execution time, and throughput while altering the workload and the number of nodes, the study [11] illustrates that HLF v1.0 surpasses v0.6 in terms of scalability, execution time, latency, and throughput. While the study demonstrates that Hyperledger Fabric (HLF) version 1.0 outperforms version 0.6 in maintaining consistent performance across various metrics regardless of network size, it neglects to explore the impact of the underlying consensus mechanism on these results. HLF v1.0 exhibits superior scalability, supporting up to 26 nodes compared to v0.6's limit of 16. However, v1.0 struggles with 10,000 concurrent transactions when exceeding six nodes. Interestingly, the evaluation reveals better performance for v1.0 with two organizations across all measures compared to a single organization setup. A crucial missing element in this analysis is a detailed examination of the specific consensus protocol employed by each version and its demonstrable influence on the reported performance metrics. This exploration would significantly strengthen the research's conclusions.

To demonstrate that the number of peer nodes in the blockchain network and the number of transactions being completed have a significant impact on HLF's performance, the study [12] performs benchmarking of HLF and points out several architectural bottlenecks in the platform, such as the orderer nodes' use of the consensus method and the network's propagation of transactions. While this study identifies bottlenecks and proposes optimizations, it doesn't delve into how the number of ordering service nodes or the choice of consensus algorithm affects Hyperledger Fabric's performance. To gain a deeper understanding of scalability and fault tolerance, future

Table 1 Summary of various studies related to performance analysis of various blockchain platforms

Ref.	Year	C1	C2	C3	C4	C5	C6	Research gaps
[21]	2017	471	Ethereum, HLF	Yes	Yes	Varying number of transactions, Concurrent transaction	Latency, Throughput, Execution time	The study only examines the platforms' execution layer, Excluding the consensus layer, which can influence performance. No detailed analysis on scalability
[24]	2017	225	Ethereum client (Parity, Geth)	Yes	No	Different clients on the same (Ethereum) platform	Processing speed	Fixed node configuration
[11]	2018	349	HLF	Yes	Yes	Number of Transactions, Number of Nodes	Latency, Execution time, Scalability, Throughput	The impact of different ordering services and varied transaction rates in the network is not investigated
[12]	2018	657	HLF	Yes	Yes	Endorsement policy, Ledger database, scalability and fault tolerance, channels and resource allocation, Block size, and Transaction arrival rate	Throughput, Latency, CPU Utilization	Doesn't explore how varying ordering service nodes or using different consensus algorithms impacts performance in Hyperledger Fabric
[13]	2018	197	HLF	Yes	Yes	Number of input, Number of chaincode, Number of peers, Number of channels	Latency, Throughput	Study to investigate the impact of multiple organizations, ordering services and number of transactions in the network
[16]	2019	242	HLF	Yes	Yes	Transaction rate, Simultaneous transactions, Number of transactions	Throughput, Latency	The study does not delve into how the varying number of peers, organizations, and nodes in ordering services influences performance metrics
[25]	2019	276	Raft consensus algorithm	Yes	No	Number of nodes, Packet loss rate, Election timeout, Number of elections	Network split time	Lack of detailed analysis of Raft's scalability in private blockchain networks with increasing node count
[17]	2020	62	HLF	Yes	No	Batch time out, Batch size, Number of endorsing peers	Network throughput, Transaction Latency	Performance evaluation is done on a considerably small network. The influence of increased transaction volume, concurrent transactions, and the number of orderers on the network performance needs to be considered
[18]	2020	49	HLF (measured with tool HLF-GLDB)	Yes	No	BlockSize, ValueSize, NumBlocks, NumWrite, NumRead, Compression, StateDBSize, Bloomfilter	Execution time, Transaction per second	Small key and single signature
[19]	2020	47	HLF	Yes	No	Transaction arrival rate, Endorsing timeout rate	Throughput, Mean transaction response time, Transaction rejection probability	An attack or hardware/software crash could cause peer breakdown
[26]	2020	230	Consensus Mechanism	Yes	No	DAG, PoS, PoW	Transaction per second, Confirmation delay, Confirmation failure probability, Average time to generate a new block	Imact of Network latency, network size on the performance of the consensus mechanisms need to be considered

Table 1 (continued)

Ref.	Year	C1	C2	C3	C4	C5	C6	Research gaps
[27]	2020	23	Hyperledger Sawtooth	Yes	No	Transaction batch, Transaction scheduler, Number of validators, Consensus mechanism	Execution time	Need to conduct a detailed analysis that considers latency and throughput
[28]	2021	46	HLF	Yes	No	Block Size, Endorsement policy, Number of organizations, Block frequency, Transport Layer Security, Programming language of smart contract, Endorser peers, Ledger database, Ordering services, Number of clients	Latency, Throughput	influence of simultaneous transactions and the impact of a varying number of transactions in the network need to be evaluated
[29]	2022	42	HLF	Yes	Yes	Private data, Crashing node, Network delay, Multiple loads	Latency, Throughput, Traffic	Need of detailed evaluation comparing the performance characteristics of Hyperledger Fabric's supported programming languages, such as Go, Node.js, and Java
[23]	2022	16	HLF	Yes	No	...	Throughput, Storage consumption, Latency, Availability	No detailed scalability assessment
[30]	2022	5	HLF	Yes	No	Transaction arrival rate, Transaction consensus rate, Consensus system capacity, Block generation rate	Response time, Throughput	No detailed scalability assessment
[31]	2022	28	Hyperledger Besu	Yes	Yes	Send rate, Load balancer, Consensus, Network size, Node flavor, Block time	Resource consumption, Scalability, Latency, Throughput	Detailed analysis on fault tolerance of consensus mechanism
[10]	2023	1	HLF, Ethereum, Hyperledger Besu	Yes	No	Different versions of HLF, Different platforms	Latency, Throughput, Success rates	Impact of varying workload and increasing the network size on the performance need to be evaluated
[32]	2023	6	HLF	Yes	Yes	TPS value, Delay Time, Number of CLI modules	Load, Amount of data, % CPU usage rate, Time	Potential bias from physical machines and limited data
Our Work	...	...	HLF	Yes	Yes	Varying transaction rate, Ordering Service, Number of organizations, Number of transactions, Programming language of smart contract	Success rate, Latency, Throughput, Resource consumption	Impact of various endorsement policies and consensus protocols with varying workloads on performance

In the table above C1, C2, C3, C4, C5, C6 are:

C1: Citation Score, *C2*: Blockchain Platform, *C3*: Performance, *C4*: Scalability, *C5*: Parameters, *C6*: Evaluation Metrics

work will involve evaluating HLF with various network structures. This will include analyzing the impact of different consensus mechanisms and the number of ordering service nodes on performance under varying workloads.

To assess the performance and scalability powers of HLF's newest version (v1.0), authors [13] have experimented with HLF on different workloads and analyzed its throughput and latency characteristics. Using a set of fabric-specific microbenchmarks, authors optimize various transaction and chaincode parameters and investigate their effect on transaction latencies. It is analyzed that for write, throughput scales nearly linearly up to a transaction rate of 1000 tx/sec. However, throughput drops significantly beyond this point, and transaction delays and latency experience a significant increase. At higher workloads, they discovered that reducing the number of necessary

endorsements (a chaincode parameter) can improve throughput and reduce latency. A stochastic reward net-based performance model for HLF v1.0+ is developed by the authors [14], and the efficacy of the model is assessed by measuring mean queue length, throughput, and utilization at each peer and pivotal processing stage within a peer Increasing block size can potentially alleviate ordering service and ledger write performance bottlenecks, but it can also introduce other trade-offs. If the transaction endorsement process is the main bottleneck for performance, parallelizing it can considerably decrease the length of the endorser queue and boost the maximum throughput of the system. While these studies provided valuable insights into performance, a key limitation remains the lack of a method to determine the optimal configuration settings for a specific use case.

Androulaki et al. [15] has investigated the impact of peer CPU, storage type (SSD versus RAM disc), and block size on blockchain throughput and latency in HLF v1.1. This study also demonstrates the system's scalability through the expansion of peer numbers. Moving forward, it is crucial to explore the utilization of benchmarks for performance optimization and evaluation and assess scalability with significantly expanded deployments. According to the paper [16], network workload has a significant impact on the performance of a blockchain in the context of scalability and latency. The findings depict that the chosen AWS EC2 instance can support up to 100,000 HLF platform participants. Shalaby et al. [17] evaluated the performance of HLF version 1.4 by examining the impact of various configurable parameters such as the number of endorsing peers, batch size, and batch timeout on network throughput and end-to-end transaction latency. Altering the batch timeout impacts both latency and throughput as the number of concurrent transactions rises. Setting the batch size to 45 transactions offers a balance between efficiency and latency. This size avoids reaching the maximum batch size before the timeout and allows all transactions to fit within a single block. However, as the number of parallel transactions within the block increases, latency tends to rise, while throughput conversely benefits. These findings have significant implications for improving blockchain performance.

The HLF-GLDB benchmark tool is developed to replicate database access in HLF. The findings illustrated in [18] show that data compression in GoLevelDB is a significant performance bottleneck in the HLF network and increased the network's overall performance by 54% by disabling data compression. Furthermore, the study discovered that database size affects performance, and it is recommended to combine smaller values to be represented by a single key to minimize database access overhead in chaincode. These findings are critical for improving the performance of HLF-based blockchain networks. To scrutinize the performance of the transaction process in HLF version 1.4, study [19] incorporated a hierarchical modeling approach based on block timeout constraints and endorsement; they created equations to estimate the system's performance. It used extensive simulations and numerical analyses to validate its model. The research is focused on two timeout constraints, which resulted in the development of a hierarchical model to describe the process of HLF transactions, beginning with the submission of transactions by clients and ending with the validation and commenting of transactions as a block is completed. The study also formulated formulas for platform throughput, mean transaction response delay, and transaction rejection probability.

In order to investigate the Fabric platform's transaction performance by examining the factors that have the most significant impact on overall overhead, the study [20] focuses on the impact of programming language and the number of endorser peers on transaction latency. Furthermore, the study compares the performance overhead of various chaincode implementations across Fabric's three programming languages: GO, Node.js, and Java. The tests are carried out with different node counts (1,2,4,6,8,10,12,14,16), and the findings show that GO outperforms the other two languages. The performance of two popular blockchain platforms, HLF and Ethereum, is compared under varying workloads [21]. The empirical findings disclose that HLF continually outperforms Ethereum throughout all evaluation metrics, including execution time, latency, and throughput.

The performance measurements of HLF version 2.0 are illustrated in the study [22] found a significant improvement in HLF version 2.0 compared to older versions of the blockchain framework. Adopting an approach that involves collecting measurement data from the system being studied to determine the pattern of resource utilization or degradation in performance that could be triggered by software aging is discussed in [23]. While this approach has been widely used in other contexts, there is limited research on the impact of software aging on blockchain platforms. Only one study has been reported in the literature, which analyzed the effects of software aging on a PoW blockchain network, specifically HLF.

2.2 Studies related to consensus

The performance complexity of distributed systems has made it difficult to fully comprehend the impact of ordering services, scalability, and bottlenecks on HLF. To address this issue, the authors [33]conducted a thorough evaluation of HLF's performance and examined traits of the execution, ordering, and validation phases. The evaluation

included a comparison of three different ordering services: Raft, Kafka, and Solo. The findings revealed that the execution phase demonstrated satisfactory scalability with the OR endorsement policy, whereas the AND endorsement policy did not perform satisfactorily. Interestingly, the authors found no substantial differences in the performance of Raft, Kafka, and Solo ordering services. However, the validation phase was noted as the probable bottleneck due to the slower speed of chaincode validation. The study has primarily concentrated on measuring the performance of consensus algorithms, neglecting to assess their resource consumption.

A novel method for evaluating the effectiveness of consensus algorithms on Ethereum and HLF private blockchain platforms is presented in the study [34]unfolds the impact of consensus techniques on the overall performance of the blockchain network by conducting a rigorous quantitative examination of important performance measures, such as latency and throughput. The results provide compelling evidence that the selection of a consensus algorithm can exert a significant constraint on a blockchain network's overall performance. Furthermore, the research demonstrates that the Practical Byzantine Fault Tolerance (PBFT) consensus algorithm consistently outperforms the Proof-of-Work (PoW) algorithm in terms of both speed and transaction processing capacity, even under diverse workloads. However, the study would benefit from further exploration of two key areas. Firstly, the impact of varying node counts within the network on these performance metrics for both platforms remains unaddressed. Secondly, the ability of these platforms to manage concurrent transactions, a crucial factor for real-world usability, is not evaluated.

An innovative approach that considers both Byzantine and network failures when evaluating the effectiveness of consensus algorithms and blockchain platforms in adversarial network conditions is presented in the study [35]. The strategy entails using stochastic modeling methods to examine how consensus algorithms behave under various ideal and unfavorable operational circumstances. After that, the effectiveness of several consensus mechanisms is tested in a blockchain application using chaos engineering techniques. Analyzing four permissioned blockchain platforms and their corresponding consensus mechanisms serves as an example of the suggested technique.

The performance metric for evaluating the effectiveness of various consensus protocols, including proof-of-activity (PoA), PoS, DPoS, and PoW based on multiple parameters such as scalability, fault-tolerance rate, latency, and degree of decentralization in the study [36] indicates that PoA and PoW protocols have low degrees of decentralization and scalability, respectively. Additionally, DPoS exhibits the highest throughput, while PoW has the highest latency and average block time. Furthermore, the study suggests that the proof-of-authority protocol is the most effective in mitigating the risk of 51% attacks compared to the other protocols. The performance of the most widely used consensus algorithm, Raft, in networks experiencing a significant packet loss rate is investigated in the study [25]. The study proposes an analytical model that accurately calculates the probability of a distributed network split occurring. The model considers the election timeout period, packet loss rate, and network size as inputs to calculate the split probability at any given time. A Raft simulator is executed to confirm the model's accuracy and demonstrate that the analytical and simulation results match. This model makes it possible to theoretically predict network split probability and optimize the parameters within the Raft consensus algorithm.

The findings of the performance evaluation of three prominent consensus mechanisms in blockchain technology - PoS, DAG, and PoW based on several metrics, such as the average time required for generating a new block, confirmation failure probability, confirmation delay, and TPS in [26] demonstrate that the consensus agreement mechanism is influenced by both network load conditions and network resources (buffer size, computation power, coin age). Moreover, the study findings reveal that PoS and PoW are more responsive to changes in network resources, whereas DAG is susceptible to variations in network load conditions.

2.3 Studies related to other blockchain platforms

HLF and Ethereum are widely acknowledged blockchain platforms that have gained significant adoption within the industry. To compare the performance of these two platforms, a study [21] conducts performance evaluation in varying workload environments and reveals that HLF consistently surpasses Ethereum in the context of throughput, latency, and execution time. Although both platforms have made progress, they still fall short of current database systems when it comes to handling high workloads. However, Ethereum is capable of handling more concurrent transactions than HLF with similar computational resources.

A comprehensive experimental analysis of Hyperledger Besu in a private blockchain setting showcases the platform's performance traits from an application viewpoint, specifically in terms of scalability, resource utilization, latency, and transaction throughput. To achieve this, the study [31] adds a load balancer middleware and investigates three proof-of-authority consensus algorithms such as QBFT, Clique, and IBFT 2.0. The evaluation is carried out using the Hyperledger Caliper benchmark tool. The

findings reveal several key insights, including the significant impact of blockchain parameters like block time and block size on Besu performance, the bottlenecking effect of transaction execution and blockchain state updates, and the scalability potential of Besu networks with QBFT consensus, which can be increased up to fourteen validators without any evident decrease in performance.

The performance of two key components in HLF blockchain systems: modular functions and state databases, is evaluated using a range of configurations and workloads to assess the impact of different combinations of modular functions and state databases on system performance, and it has been observed from the study [32], that the performance of the state database is closely correlated with read-type transactions in HLF blockchain systems, while it has little to no impact on write-type transactions. This suggests that the choice of state database should be carefully considered, especially for applications that involve a high volume of read-type transactions.

In comparison to existing studies, our research offers a deeper understanding by focusing on a specific platform, Hyperledger Fabric. This research conducts an exhaustive exploration of the platform's performance characteristics under a variety of operational configurations Table 1. Our work investigates the influence of key factors on scalability and throughput, including transaction volume, ordering service, the number of participating organizations, and the programming language employed for smart contract development Table 2. This work focuses on Hyperledger Fabric. An in-depth examination of various factors provides valuable and actionable insights for optimizing its use in real-world deployments, addressing a gap in existing research primarily focused on broader architectural frameworks.

3 Tools and techniques

HLF offers a broad range of customizable features and potential applications for developers to establish a bespoke blockchain business network comprising heterogeneous clients with distinct roles assigned to them. Additionally, it facilitates the creation, deployment, and execution of chaincodes/smart contracts on the platform. The methodology adopted to conduct the experiment is shown below in Fig. 3.

To assess and quantify the effectiveness of the HLF platform, the utilization of Hyperledger Caliper is implemented (Fig. 4), which can produce a performance report tailored to specific use cases.

3.1 Hyperledger fabric

Hyperledger is a collaborative open-source project that originated from the Linux Foundation, intending to promote the development of blockchain technology across various industries. It involves the participation of leaders from different sectors worldwide, and its goal is to advance the application of blockchain collaboratively and innovatively. An open-source business application offers significant benefits, including a great deal of flexibility due to the availability of open-source code, modular components, and adherence to industry standards. This permits organizations to customize the technology to meet their special necessities and achieve maximum usability with minimal effort. HLF is a blockchain project under the Hyperledger umbrella that utilizes smart contracts and ledger to enable transaction management among participants. However, it differs from other blockchain systems in that it is permissioned and private. On the contrary, to open permissionless systems that allow unknown entities to join the network, HLF requires participants to conscript through an entrusted Membership Service Provider (MSP) to ensure the network's security. This means that it does not rely on protocols like "PoW" to validate transactions and maintain network integrity. To ensure transaction privacy within the HLF network, a system called a channel is utilized to isolate transactions and restrict access only to authorized nodes within that specific channel. The various roles concerned with the flow of transactions within a channel are described below.

3.1.1 Peer

Peers are a crucial component of the HLF network as they serve as hosts for ledger and chaincode instances. In blockchain technology, specifically on the Hyperledger Fabric platform, a chaincode refers to a smart contract. It's a small program that runs on the blockchain network and dictates the rules for transactions. These programs are written in languages like Go, Java, or Node.js and manage the data stored on the blockchain. Essentially, they automate and enforce agreements between different parties involved in a blockchain transaction. Essentially, chaincode acts as the business logic layer of a blockchain application, specifying how assets can be created, transferred, and updated on the ledger. Each peer has the capability to install multiple chaincodes and ledgers. While each peer doesn't need to install chaincode, peers can communicate with one another through the gossip protocol to update and modify the latest state of the ledger. This allows for efficient and decentralized communication among peers within the network. The organization peers

Table 2 Configurable parameter for HLF performance analysis

Examination	Parameters	Transaction rate (TPS)	Number of transactions (TXNs)
Impact of transaction rate TPS	Varying TPS and keeping TXNs constant	25,50,75,100,125,150,175,200,225,250,275,300,325,350	1000
Impact of TXNs (number of transactions)	Varying TXNs and keeping TPS constant	200	10, 100, 1000, 10,000, 100,000
Impact of ordering service	Varying TPS and keeping TXNs constant	25,50,75,100,125,150,175,200,225,250,275,300,325,350	1000
	Varying TXNs and keeping TPS constant	200	10, 100, 1000, 10,000, 100,000
Impact of programming language	Chaincode written in Go language also keeping TPS constant while varying TXNs	200	10, 100, 1000, 10,000, 100,000
	Chaincode written in Go language also keeping TXNs constant while varying TPS	50, 100, 150, 200, 250, 300,350	1000
	Chaincode written in Node.js also keeping TPS constant while varying TXNs	200	10, 100, 1000, 10,000, 100,000
	Chaincode written in Node.js also keeping TXNs constant while varying TPS	50, 100, 150, 200, 250, 300,350	1000
Impact of multiple organizations	Increased number of organizations with varied TPS and constant TXNs	25,50,75,100,125,150,175,200,225,250,275,300,325,350	1000
	Increased number of organizations with constant TPS and varied TXNs	200	10, 100, 1000, 10,000, 100,000

within the HLF network can be categorized into several types: leader, anchor, committing, and endorsing peers. Each peer is assigned specific behaviors and operations to carry out within the network. The role of a leader peer node is crucial in the network because it receives newly generated blocks from the orderer and distributes them to other peers through the gossip protocol. On the other hand, anchor peers are known to other organizations within the network, allowing them to communicate with each other using the gossip protocol. Endorsing peers are responsible for endorsing transactions that are sent to them by the ordering services. These peers play a vital role in validating the transactions before they are committed to the ledger. In contrast, committing peers do not execute transactions but receive the transactions from the orderer and commit those transactions to their local copy of the ledger. This ensures that the network maintains a consistent and accurate record of all transactions.

3.1.2 Orderer

The HLF network relies on the orderer to pack the transactions into blocks and send them to anchor peers across the network. The flow of transactions within the network consists of three distinct stages: The proposal stage, the packaging stage, and the validation stage. Although the orderer is primarily responsible for the packaging step, it also contributes to the validation step by sharing new blocks with the network. By doing so, the orderer ensures that the network maintains a precise and current ledger of all transactions. In the HLF network, the orderers package transactions into blocks and send them across the network and play a vital role in enforcing basic access control for channels. They control who has the authority to read and write data to the channels and configure them. It is important to note that the ability to modify configuration elements within a channel is subject to policies established by the relevant administrators during the creation of the consortium or the channel. By enforcing access control and adhering to established policies, orderers ensure that the network operates in a secure and efficient manner. The Fabric 2.0 version recommends the utilization of Raft ordering services, which have replaced the previously deprecated Kafka ordering services.

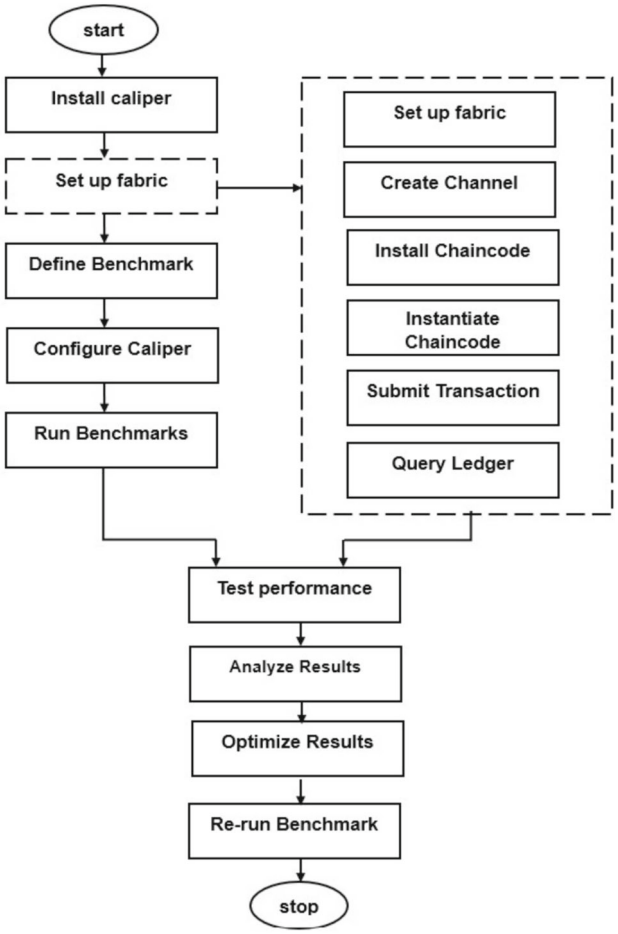


Fig. 3 Methodology employed in the experiment

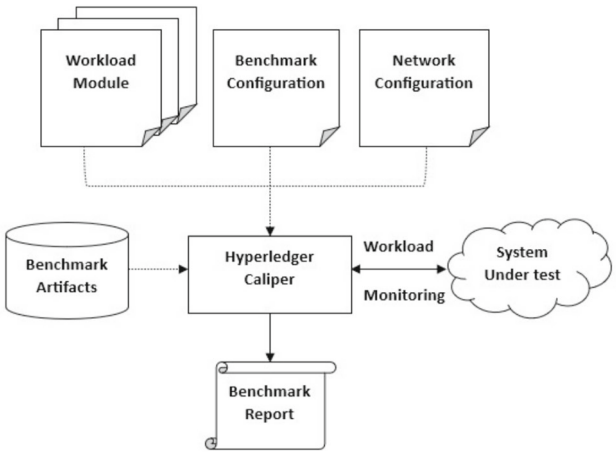


Fig. 4 Implementation of caliper for executing benchmark

3.1.3 Organization

An organization in HLF refers to a group of peers that work together towards a common goal in conducting business. These peers can be made up of one or more members who collaborate to maintain and operate the network. Each organization has its own MSP which oversees the identity and access control of its members. By dividing the network into separate organizations, HLF allows participants to transact and collaborate securely while still retaining control over their own data and operations.

3.1.4 Failed transactions

Blockchain transactions may get rejected or fail to be processed due to a variety of factors, including errors in syntax, errors in achieving consensus among nodes, and errors arising from the version of the blockchain software being employed. The consensus errors in blockchain occur when there is a dispute between the nodes in the network about the validity of a transaction. This can transpire due to various reasons, such as a network partition or a malicious node trying to manipulate the consensus [37]. In the case of HLF, validation logic errors can occur when the transaction's validation system endorsement policy is not satisfied and the peer nodes cannot reach a consensus on the validity of the transaction. This endorsement policy is defined by the chaincode and specifies the number of required endorsements, the identities of the endorsing peers, and the criteria for successful endorsement. If the policy is not realized, the transaction is not endorsed and will be rejected. For instance, in HLF, the endorsement policy specifies the required amount of signatures from endorsing peers who must agree that the transaction is valid. If the endorsement policy is not met, the transaction is deemed invalid and is rejected by the system. The policy ensures that transactions completed on the blockchain are valid and fulfill the network's predetermined requirements. Our experiment observed transaction failures related to endorsement policy, conflicts, and timeout. There are various reasons outlined for transactions to fail in the HLF blockchain, some of which are described below:

Insufficient Endorsement: Before a transaction can be added to the ledger, it needs endorsement from a certain number of peers, and not meeting the requirement of a specified number of endorsements leads to transaction failure.

Inadequate Resources: Due to the lack of resources available, such as processing power and network bandwidth, transactions may fail in HLF since it depends on the network of peers to validate transactions.

Versioning Conflict: Multiple versions of smart contract/chaincode deployment on the network are allowed by the HLF blockchain, but if the transaction submitted to the smart contract is not compatible with the network's version, then the transaction leads to failure.

Transaction Validation Failure: If the validation mechanism in the HLF smart contract is poorly described,

a validation fault results in transaction failure and unsuccessful transaction validation.

Network Congestion: Hyperledger blockchain network communicates in a peer-to-peer manner to reach consensus, and if the network is congested or encounters latency issues, the transaction in the network fails too.

3.2 Performance and scalability

Analyzing the performance of blockchain platforms is pressing in providing the guarantee to the users that it can help in delivering the required quality of service and meet the demands of its deliberated use cases. The user can distinguish limitations and bottlenecks in the scalability and performance of the system through performance analysis. The performance analysis can serve in optimizing the resource utilization of the system, such as network bandwidth, CPU, and storage, in order to enhance the overall efficiency [38]. Furthermore, performance analysis can assist in determining the possible security concerns as well as ensuring the system can manage high transaction volume without jeopardizing its integrity and security. This research aims to establish a framework for evaluating the performance of blockchain technology. The evaluation will encompass key metrics such as success rate, throughput, latency, and resource consumption (CPU and memory). The scalability of the HLF can be evaluated with the varying workloads, number of transactions, organization, and ordering services. Scalability aims to increase the network's capacity to process significant transactions without compromising its security and decentralized structure.

3.2.1 Throughput

Throughput under the frame of blockchain depicts the number of transactions that are successfully processed by the network within a specific time frame, typically measured in seconds. The throughput performance metric indicates the speed and efficiency of the blockchain network in processing transactions. Read throughput is a way to evaluate performance that measures how many successful read operations occur during a certain timeframe. The metric is typically expressed as reads per second and provides a specific count of completed read operations within the given interval [37].

Read Throughput = Total read operations / Total time in seconds

Transaction throughput is a measure of the speed at which valid transactions are confirmed and added to the blockchain during a specific timeframe, typically expressed as TPS. This performance metric evaluates the overall network's efficiency, taking into account all nodes rather than just a single node.

Transaction Throughput= Total valid committed transactions / Total time in seconds

By increasing the throughput, the network can manage more transactions, thereby improving the entire efficiency and scalability of the blockchain network.

3.2.2 Latency

The latency of a transaction is calculated by finding the difference between the time taken to complete the transaction and the time taken to deploy it. Latency is divided into two categories [37], transaction latency and read latency. Read latency is determined as the time between the response of the request and the submission of the read request.

Read Latency = Reply received time - Request submission time

On the other hand, the amount of time needed for a transaction to become available throughout the entire network is stated as transaction latency. This period encompasses the interval from the moment the transaction is first submitted and ends up being widely accessible on the network.

$$Transaction\ \ Latency = confirmation\ \ time \times network\ threshold - transaction\ \ submission\ \ time$$

The average latency measured in our work states the mean latency of all the transactions in a given data set.

3.2.3 Resource Consumption

During the test execution, transactions utilize the resources such as CPU, cache memory, disk, and RAM. Extreme consumption of these resources, e.g., disks, can induce a significant impact on the systems' performance [23]. The Resource Monitor allows initiation/termination of monitoring and obtaining information about the resource usage of the underlying blockchain system, such as network I/O, memory, CPU, and so on. The Docker resource monitor provides access to several performance metrics, including maximum and average memory usage, maximum and average CPU usage as a percentage, inbound and outbound network traffic, disk read operations, and disk write operations. This information can be used to monitor and manage the resource consumption of Docker containers, allowing for efficient use of system resources and improved performance.

In network I/O, traffic in and traffic out refer to the amount of data that is being received by a peer or orderer node (traffic in) and the amount of data that is being sent out from the node to other network participants (traffic out). The figures depicting the data on network input/output

(I/O) traffic, measured in megabytes (MB), are denoted in Fig. 5. The evaluation is performed with varying numbers of peers and transaction rates while keeping the transaction volume constant at 1,000 transactions. The resultant graph reveals a notable trend that there is a demonstrable rise in average inbound traffic as the number of peers within the network increases. High Traffic In and Traffic Out indicate significant data transfer between the client and the blockchain network. This can help assess the network's capacity to handle data volume. Comparing Traffic In and Traffic Out for different workloads can reveal which scenarios generate more data transfer. This aids in understanding the impact of various use cases on the network.

Several use cases are outlined by varying the key parameters, such as transaction rate and number of transactions, in order to perform the test on the Hyperledger Caliper. These use cases enable assessment of the system's performance and functionality under different situations and conditions, which can help identify potential bottlenecks and improve system efficiency.

3.2.4 Success rate

The success rate of a blockchain can be measured by evaluating the ratio of successfully executed transactions to the total number of transactions attempted. Successful transactions refer to those transactions that are completed without encountering any errors or exceptions during their execution.

4 Experimental results and analysis

The subsequent section demonstrates the different configurable parameters cataloged in Table 2 along with the experimental setup specified in Table 3 to conduct the evaluation. The experiment involves utilizing a personal computer equipped with advanced hardware specifications, including an AMD Ryzen 5 5500U processor with Radeon Graphics, 16GB of RAM, and a 512GB SSD, running the Ubuntu operating system version 22.04 LTS. The performance of the HLF version 2.4.9 platform is assessed through rigorous testing using the Hyperledger Caliper tool [39], including versions v0.5.0. The impact of varying the transaction rate (rateControl), varying the number of transactions, the programming language adopted to write the smart contract, and the number of organizations on the blockchain's performance. A simple money transfer chaincode that incorporates the function open() and query() is employed in the experiment. With the use of the open function, a new account is generated and initialized with the specified amount of money.

The query function retrieves the amount of money in a given account. It has been observed that the complexity of the chaincode used to execute transactions can also impact performance. More complex chaincode can require more processing power, leading to slower transaction processing times. As established earlier, this study involves varying parameters such as transaction rate, with rates ranging from 25TPS to 350TPS while keeping the number of transactions fixed at 1000. Additionally, the number of transactions is varied while keeping the transaction rate fixed to observe the significant changes in throughput and latency, the resource consumption of the various components, and identify potential areas for optimization in the system. By varying these parameters and observing the resulting changes in performance, this study aims to obtain a more exhaustive understanding of the behavior of the system under different scenarios and conditions.

The performance of a blockchain is not only affected by transaction rate and number but also by the number of peers, orderers, and organizations. In this study, the number of organizations is varied from one to three with the focus on observing the impact on the performance of the

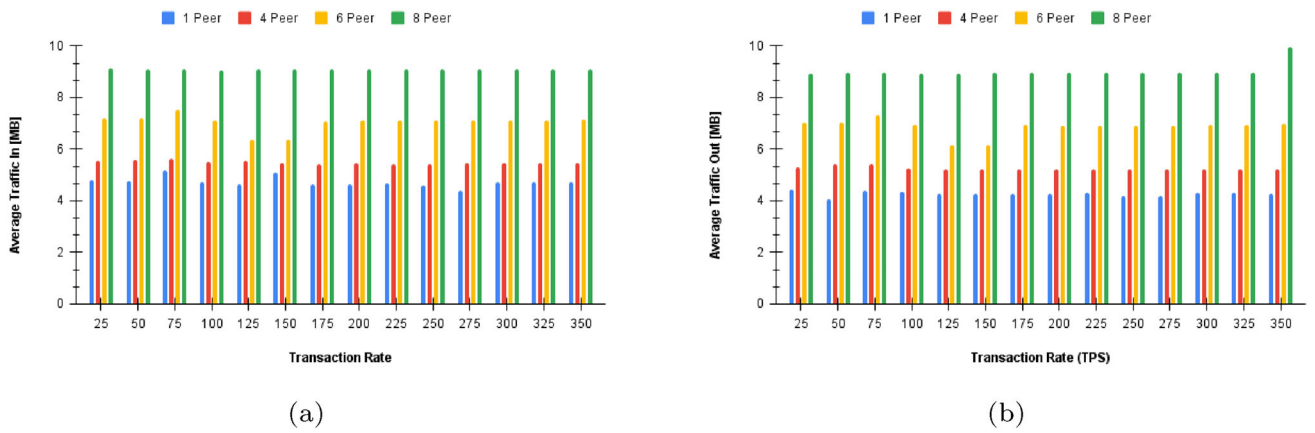


Fig. 5 Graphical representation of resource monitor of traffic in [MB] and traffic out [MB] with multiple peers

Table 3 Experimental setup and configuration for hyperledger fabric performance evaluation

Component	Description	Parameters
Platform	Hyperledger Fabric	v2.4.9
Number of Organizations	The number of independent organizations participating in the network	1, 2, 3
Database	State Database	CouchDB
Channel	Private communication pathway	Single private channel
Number of Ordering Service	The number of nodes participating in the ordering service (Solo, Raft)	1, 3, 5 orderer
Chaincode Implementation	The programming language used to develop the chaincode (smart contract)	GO, Node.js
Transaction Type	The type of transaction being executed	Open, Query
Transaction Rate	The number of transactions submitted per second to the network	25 TPS, 50 TPS, 75 TPS, 100 TPS, 125 TPS, 150 TPS, 175 TPS, 200 TPS, 225 TPS, 250 TPS, 275 TPS, 300 TPS, 325 TPS,350 TPS
Number of Transactions	The number of transactions submitted	10, 100, 1000, 10,000, 100,000

blockchain. The experiment employs a multi-round test case where the transaction rate is progressively increased in each round. Furthermore, each new round has a higher transaction rate than the previous round. We have conducted each experiment at least 20 times for each defined parameter.

4.1 Impact of transaction rate on latency and throughput

Figure 6a illustrates the throughput and average latency for 1 organization with 1 peer and single orderer over different transaction rates, and Fig. 6b illustrates the throughput and average latency for 2 organizations with 2 peers (1 peer per organization) and single orderer over different transaction rates as well.

4.1.1 Observation

As mentioned previously, the chaincode deployed in the Fabric network is related to a simple money transfer between the accounts and has the function open and query. Throughput is evaluated by varying transaction rates with each round, starting from 25 transactions per second (TPS) and increasing by 25 TPS in each subsequent round until reaching 350 TPS. From Fig. 6 (a), it can be noticed that for open( ), transaction rate ranging from 25 TPS to 150 TPS, there is a linear growth in throughput, and the throughput delivered is 25 TPS, 49.9 TPS, 74.5 TPS, 99.7 TPS, 105.7 TPS, and 119.6 TPS respectively. However, After the transaction rate of 150 TPS, throughput started to decrease as the transaction rate increased. This suggests that the system may encounter bottlenecks as the load increases beyond a certain point. Based on the result, it can

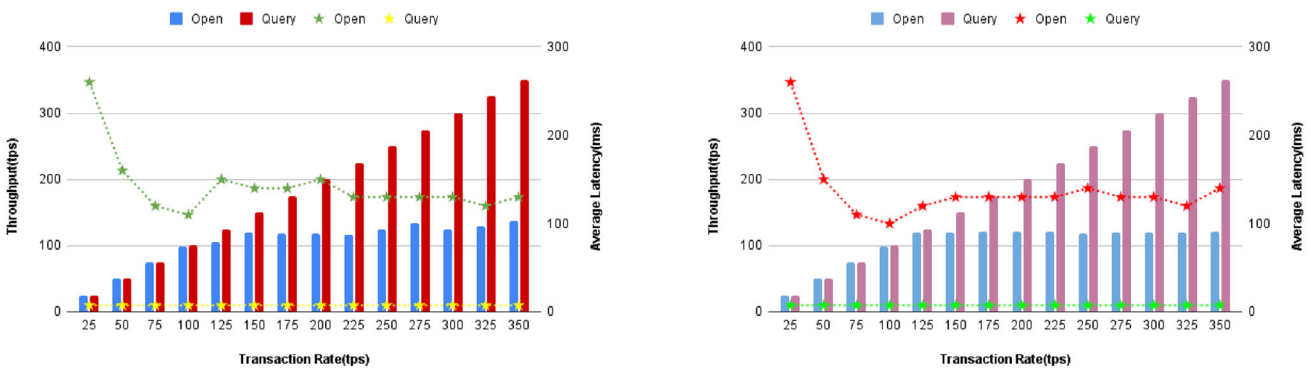


Fig. 6 Impact of transaction rate on latency and throughput (**a**) 1 organization 1 peer and (**b**) 2 organizations 2 peers

be inferred that the test environment used in the study is capable of handling open() transactions up to a rate of 150 transactions per second (TPS). At a transaction rate of 25 TPS, the average transaction latency is higher than rates ranging from 50 TPS to 350 TPS. Specifically, the average latency at a rate of 25 TPS is around 0.26 s or 260 milliseconds.

According to the results presented for our test environment, the average latency of transactions is at its minimum when the transaction rate is 100 TPS, with an average latency of 110 milliseconds or 0.11 s. After the transaction rate of 225 TPS, the average latency varies between 130 ms to 120 ms. This indicates that the system is effective when handling transactions at a 100 TPS rate and that increasing or decreasing the transaction rate may result in higher latencies and can impact performance. Likewise in Fig. 6b, for open (), it can be seen that from transaction rate 25 TPS to 200 TPS, there is an increase in throughput (TPS) as compared to (a). Moreover, the average latency is relatively low in Fig. 6b until the transaction rate of 175 TPS is reached, after which there is some variation in the average latency between 130 ms to 140 ms respectively. It has been observed that in this test environment either one or two organizations, the query type transaction can be executed at a rate of 350 transactions per second (TPS) without experiencing any significant delays.

4.2 Number of transactions' impact on latency and throughput

Figure 7 demonstrates the impact on throughput and average latency by varying the number of transactions over a fixed transaction rate. The broad array of transactions 10,100,1000,10,000,100,000 are submitted to the blockchain at the fixed rate of 200 TPS. Figure 7a illustrates the impact on throughput and latency for 1 organization with 1 peer and (b) depicts the impact for 2 organizations with 2 peers (1 peer per organization) by varying the workload.

4.2.1 Observation

Figures 7a and 6b evaluate the influence of the number of transactions on average latency and throughput. For open(), 10 transactions are sent at the rate of 200 TPS (as shown in Fig. 7a), the observed throughput is 46.5 TPS, and the average latency is high, measuring 180 ms or 0.18 s. Similarly, in Fig. 7b, throughput is observed to be 39.4 TPS, and the average latency is 210 ms or 0.21 sec. In addition, the throughput for 100, 1000, 10,000, and 100,000 transactions for 1 organization and 2 organizations are 110.7, 117.5, 134.3, 135 and 111.2, 121.8, 135.5, and 135.2 TPS, respectively. Figure 7a and b illustrate the impact on latency and throughput when 10 transactions are sent at a rate of 200 TPS in query type transaction, and the observed latency in both cases is 0.02 sec, which is slightly higher than the average latency observed for 100, 1000, 10,000, and 100,000 transactions. This indicates that the blockchain system's performance is impacted by the number of transactions being processed. It has been observed that up to 100,000 transactions are supported by this system depending on the function (open, query) with very minimal latency.

4.3 Impact of ordering service on latency and throughput

The Figures presented demonstrate the impact of ordering services on average latency and throughput. Figure 8 (a) illustrates the impact on throughput by employing the Raft ordering service in the 1 organization 1peer, while Fig. 8b illustrates the same for the 2 organizations 2 peers scenario. Graphical representations Fig. 8c–e showcase the variation in throughput and latency exerting the Solo ordering service in 1 organization 1 peer (case 1), 2 organization 2peer (case 2), and Raft ordering service in 1 organization 1peer (case 1) and 2 organization 2peer (case 2).

4.3.1 Observation

It can be observed from Fig. 8a for open() that there is a linear growth in throughput with the varying range of transaction rate from 25 TPS to 150 TPS and again shows a rise in throughput can be seen at 200 TPS. On the other hand, the observed average latency in the same scenario is higher at 25 TPS and 50 TPS as compared to the transaction rate ranging from 175 TPS to 350 TPS. In contrast to the results shown in Fig. 8a, the average latency observed for transaction rates ranging from 75 TPS to 350 TPS is between 0.12 s to 0.11 s, indicating a lower latency compared to the results obtained in (b). The (c), (d), and (e) figures describe the impact on throughput and latency by employing Solo and Raft ordering services. It can be observed that the Raft ordering service handles up to 100,000 transactions very effectively with minimum delay. The results obtained from our test environment show that the Raft ordering service performs effectively in the context of increased throughput and less latency as compared to the Solo ordering service in both cases i.e. with varying transaction rates and varying numbers of transactions.

4.4 Impact of programming language on latency and throughput

In the HLF blockchain, a chaincode is a program or a smart contract that establishes the business logic and rules governing the interactions between participants in a blockchain

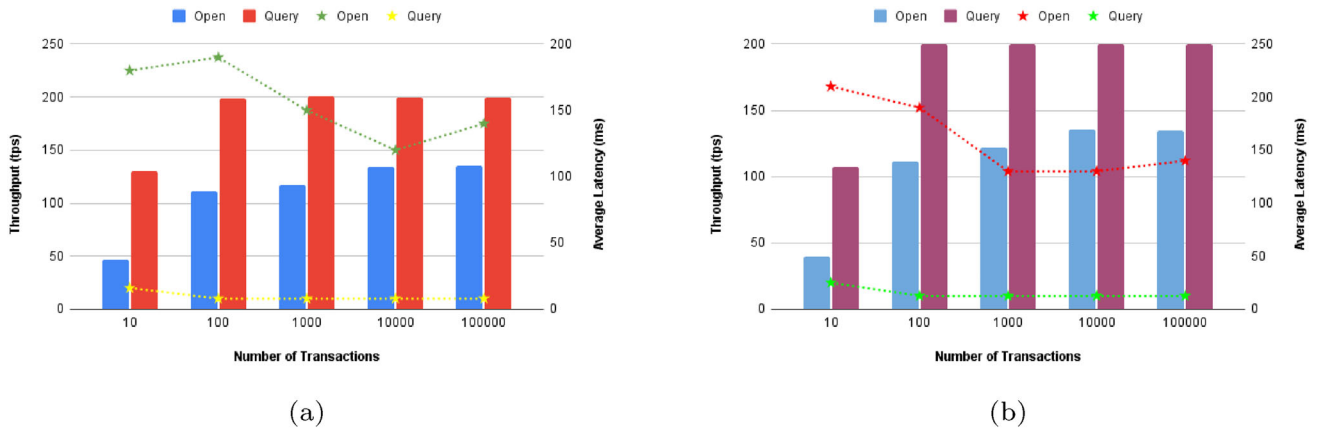


Fig. 7 Number of transactions impacting the performance from the perspective of latency and throughput (**a**) 1 organization 1 peer (**b**) 2 organization 2 peers

network. Chaincode is deployed on a specific channel of a blockchain network and is utilized to read and write the state of the ledger. Chaincode in the HLF blockchain can be written in Node.js or Go language. The chaincode employed in this study is written in Go language and Node.js and evaluated to see the impact on latency and throughput under different transaction rates (Fig. 9a, b) and the number of transactions (Fig. 9c, d).

4.4.1 Observation

Figure 9a depicts the performance of a smart contract written in the Go language while (b) shows the performance of a smart contract written in Node.js. In Fig. 9a, b, for open() type transactions, there is a linear growth in throughput as the transaction rate increases from 25 TPS up to 200 TPS implying that the system is capable of efficiently processing transactions up to a rate of 200 TPS. According to the results, the chaincode developed in the Go language outperforms the one developed in Node.js in the context of throughput and average latency, however, the difference is not very significant. The Go language-based chaincode shows slightly lower latency and higher throughput compared to the Node.js-based chaincode. Go language is considered more efficient for chaincode development in HLF due to its static typing, compiled nature, and optimized runtime performance. Node.js, on the other hand, is dynamically typed and interpreted, which may result in lower performance and scalability for complex chaincode applications.

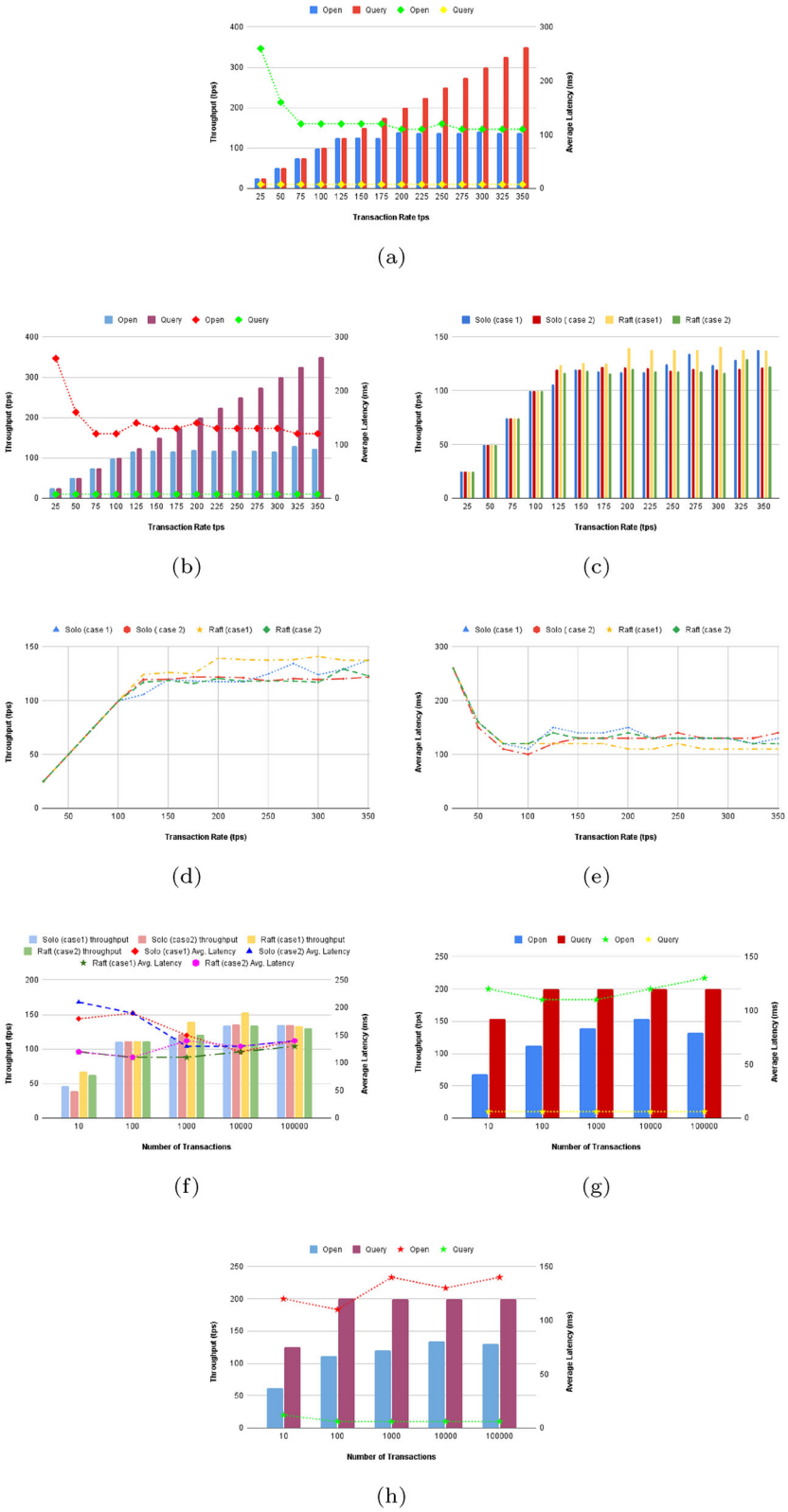
4.5 Impact of multiple organizations on Latency and Throughput

In this experiment, 3 organizations with one peer in every organization are observed in terms of throughput and latency. Figure 10a and b depict the variation in throughput and latency as the number of organizations increases. Additionally, Fig. 10c and d present the findings with respect to throughput and latency for different numbers of transactions across multiple organizations. Figure 10a and c employ Solo ordering services, while Fig. 10b and d employ Raft ordering services.

4.5.1 Observation

Graphs illustrate that the inclusion of a third organization results in linear growth in throughput for transaction rates ranging from 25 to 175 TPS and is also able to handle 100,000 transactions with a small delay. The results show that the inclusion of a third organization has led to a decrease in throughput compared to the scenarios with a single organization and two organizations. Additionally, after reaching 175 TPS, there is no significant change in the throughput observed. Figure 10a, b, and d show an increase in the average latency, indicating that the inclusion of a third organization in the network results in higher latency. It can be implied from the results that adding more organizations to a network, can potentially result in increased complexity and communication overhead, which may negatively impact the network's performance in terms of latency and throughput.

When evaluating the performance of a blockchain network using the success rate as a measure, notable findings indicate that both the open and query operations have consistently achieved a 100% success rate across all evaluation parameters. This indicates that the permissioned blockchain, under analysis, has consistently executed both open() and query() without encountering any failures or errors.Based on the observations from Fig. 11, it has been noticed that the CPU resource utilization during the execution of different workloads shows that in case 1(Using



Solo ordering service with 1 organization and 1 peer), case 2 (Using Raft ordering service with 1 organization and 1 peer), case 3(Using Solo ordering service with 2 organization with 1 peer per organization), and case 4 (using Raft ordering service with 2 organizations with 1 peer per organization), as well as case 5 (using the Go

Fig. 8 Impact of ordering service on throughput and average latency

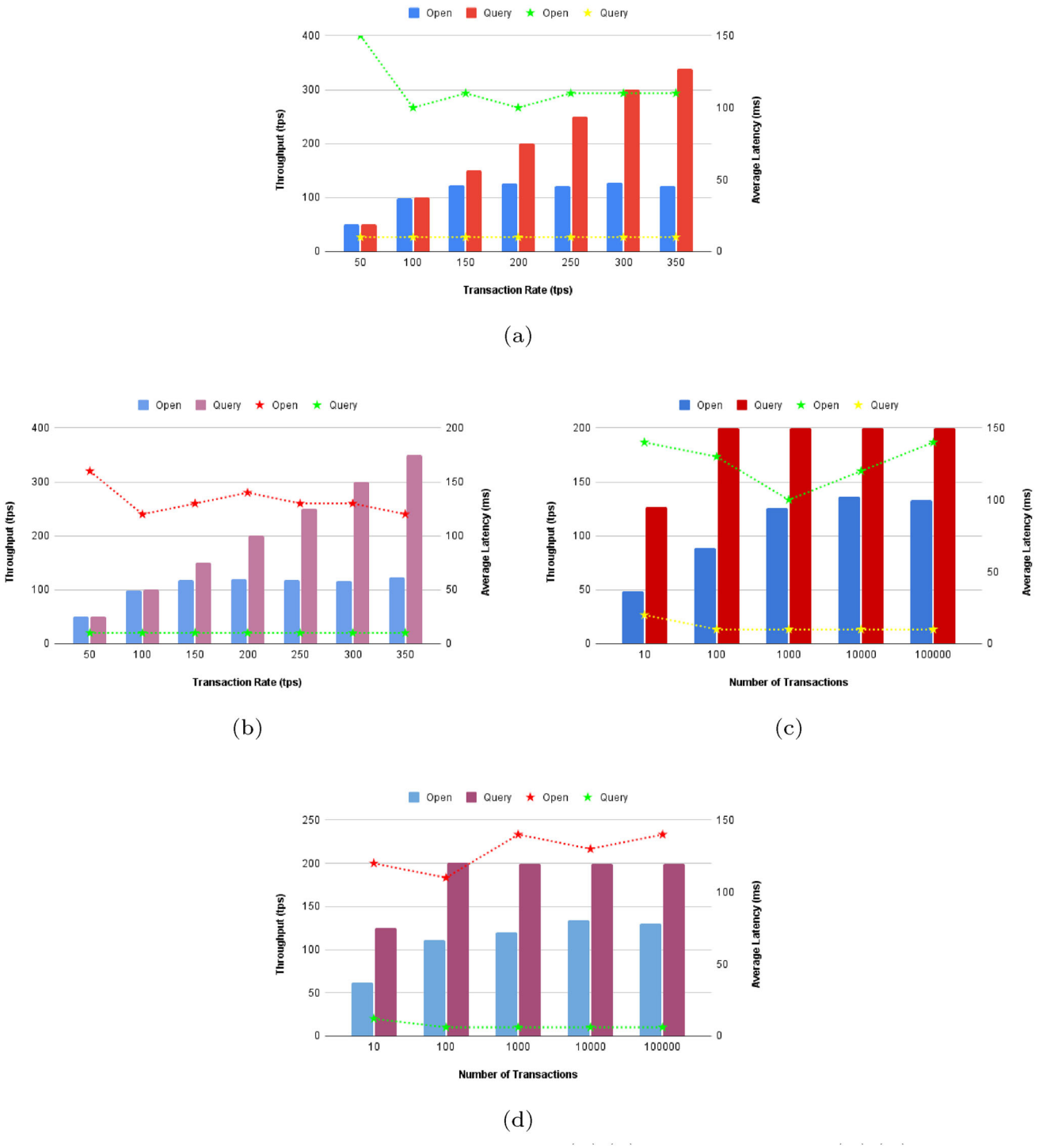


Fig. 9 Chaincode written in go language (**a**, **c**) and Node.js (**b**, **d**) to evaluate latency and throughput experienced under different transaction rate and the number of transactions

language to write chaincode), and case 6 and case 7 (using 3 organizations with 1 peer per organization), the CPU usage remains below 50%. In simpler terms, the CPU usage for these scenarios, including various configurations and setups, is relatively low and does not exceed half of the available processing power. This indicates that the executed workloads do not heavily strain the CPU resources, suggesting that the system has ample capacity to handle the workload without reaching its maximum processing capability.

HLF is a widely adopted and compelling choice for blockchain-based applications due to its modular

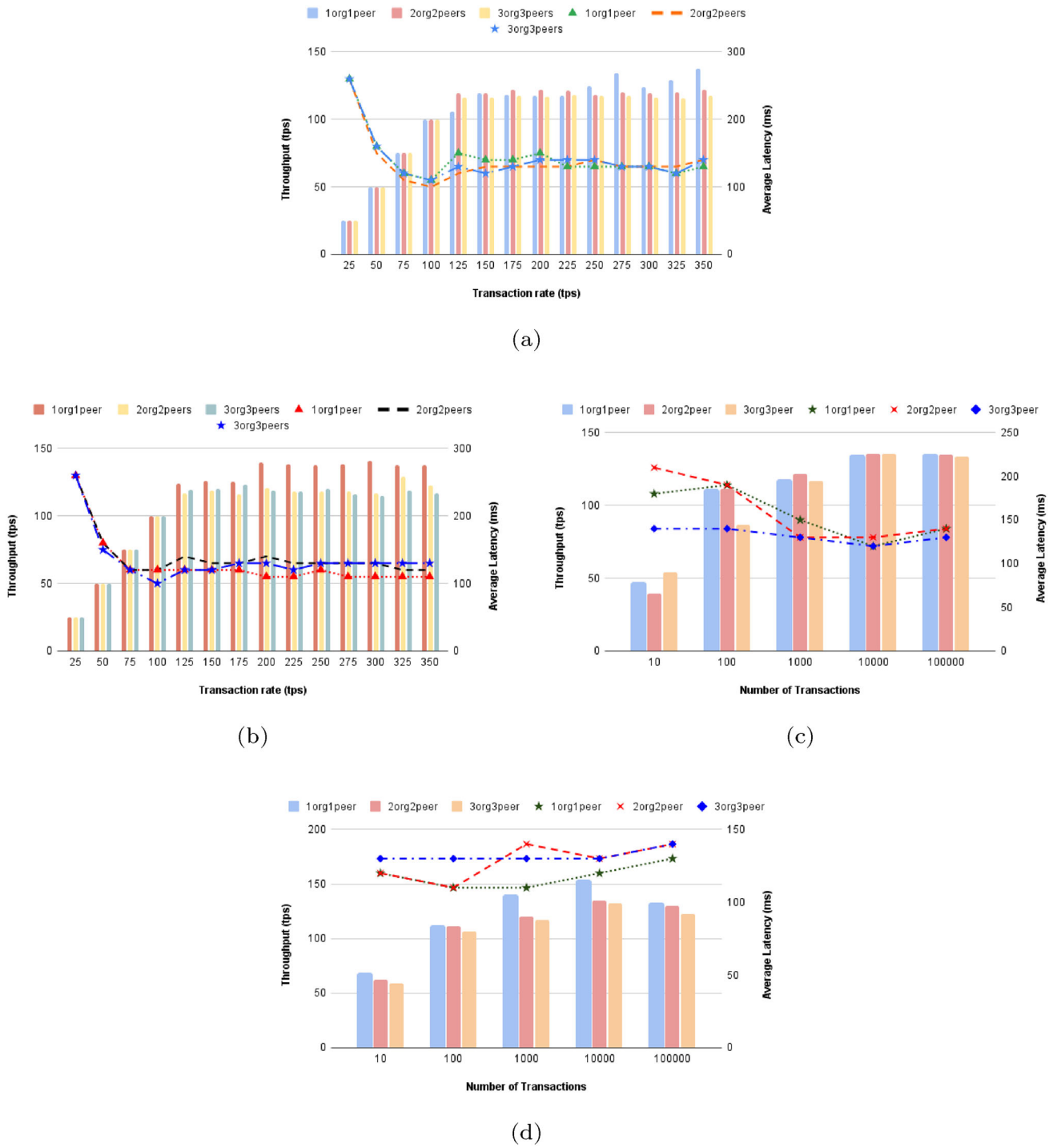


Fig. 10 Impact of the increased number of organizations on latency and throughput experienced under different transaction rates and the number of transactions

architecture, support for confidential transactions, and scalability features. HLF has its shortcomings. Hyperledger Fabric uses Optimistic Concurrency Control (OCC), which means it simulates transactions on individual peers before endorsement. This can lead to failures if multiple transactions try to modify the same data concurrently. Also, some failures that occur due to invalid signatures, mismatched world states, or policy misconfigurations can hinder successful transactions. In our study, we noticed instances of transaction failure, including occurrences such as MVCC_READ_CONFLICT, ENDORSEMENT_POLICY_FAILURE, and the timeout expired.

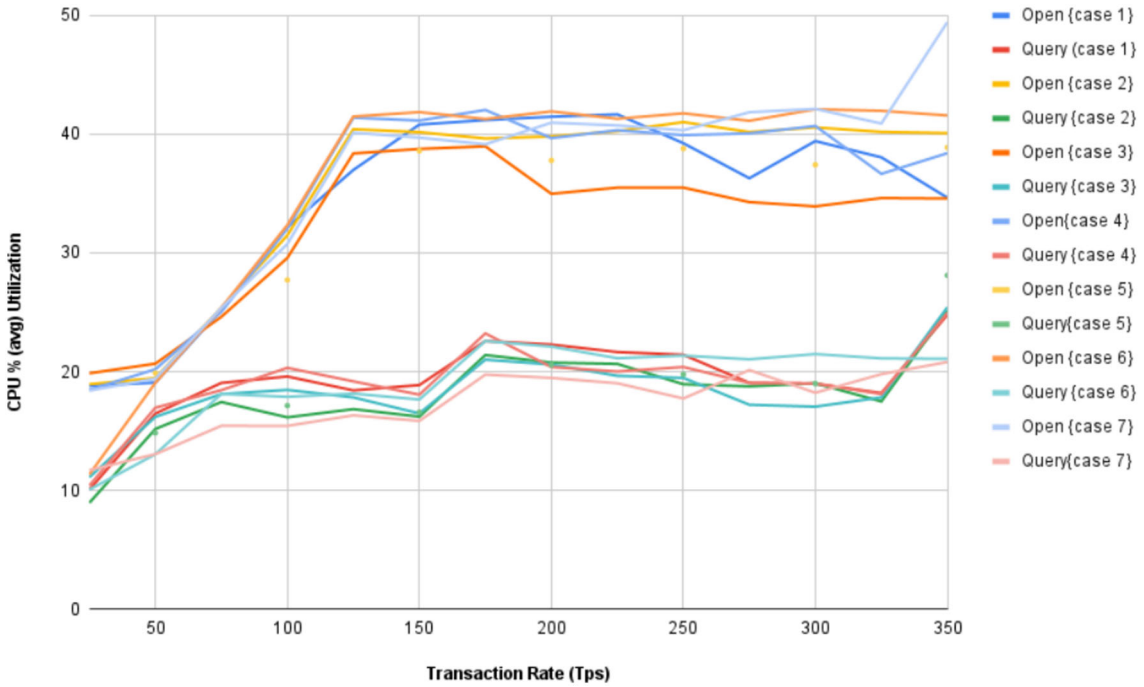


Fig. 11 CPU% average utilization in executing open() and query() in a varied number of transaction rates for different numbers of organization and ordering services

The read-write conflict arises when multiple transactions try to modify the same data concurrently. Let Txn_i and Txn_j be transactions, and their respective read sets and write sets after endorsement simulation are represented as $Read(Txn_i)$: Read a set of transaction Txn_i. $Write(Txn_i)$: Write set of transaction Txn_i. Assuming Txn_i is ordered before Txn_j within the same block (i.e., Txn_i is confirmed before Txn_j), then Txn_j is considered a conflicting transaction and will be marked invalid during validation if any of the read-write conflict or write-write conflict conditions hold. HLF utilizes the Multi-Version Concurrency Control (MVCC) mechanism to manage the consistency of records. This involves tracking changes to a record's version through a structured format consisting of key-value pairs along with corresponding version information, e.g., key: Value, version.

As shown in Fig. 12, when the first transaction reaches the validation stage, MVCC detects a mismatch. The version of the key it read earlier doesn't match the current version (due to the second transaction's modification). This mismatch triggers the MVCC_READ_CONFLICT error. In HLF, transactions require endorsement from peers before being committed. The endorsement policy defines which organizations or specific peers must endorse the transaction for it to be valid. There could be various reasons for the endorsement policy failure, such as missing endorsement, incorrect chaincode implementation, and read-write set mismatches, as concurrent writes can cause version inconsistencies leading to endorsement failure. To resolve these failures, it needs to ensure that the policy aligns with the transaction's participants, utilize Caliper's features to verify endorsement targets, and examine chaincode logs for errors.

5 Discussion

By fostering trust in a decentralized environment, blockchain technology underpins its core functionalities, including robust security, data privacy, scalability, and efficient performance. In private deployments, Hyperledger Fabric frequently surpasses Ethereum (Private Deployment) and Besu (Ethereum's Client) when it comes to

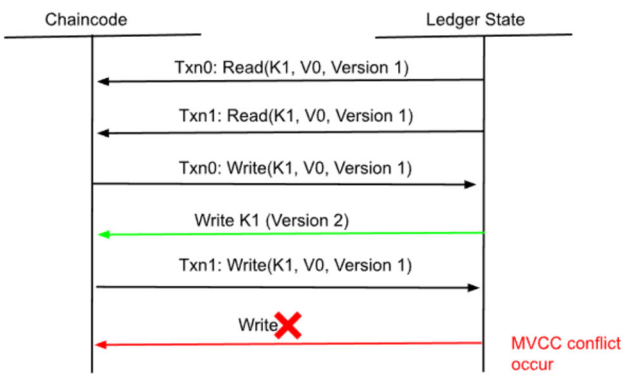


Fig. 12 An example to illustrate read-write conflict

throughput and latency, as illustrated in Fig. 13. This research is more focused on analyzing the performance characteristics of HLF under varying workloads. This section discusses the findings of this research work and delves into the reasons behind the observed behavior.

Reasons for the Observed Behaviors: Over the course of the study, we have analyzed the factors that may influence the performance of the HLF blockchain. It has been observed that the number of organizations, orderers, and peers has a substantial impact on performance and scalability. A network with too many peers and organizations can result in slower transaction processing times, whereas a network with too few peers can result in a lack of fault tolerance. The performance of the network is influenced by both the transaction rate and the number of transactions. As the transaction rate increases, there is a linear growth in throughput until a certain point. Our deployment of Hyperledger Fabric encountered a critical performance bottleneck during a recent load test. As we scaled transaction volume beyond one million, a significant number of transactions began failing with a "timeout expired while executing transaction" error. This behavior suggests that the network's capacity to process transactions was exceeded, leading to timeouts. Similarly, increasing the transactions per second (TPS) beyond a certain threshold resulted in the same error. HLF offers different ordering services, such as Solo, Kafka, and Raft, each having its own strengths and weaknesses and features depicted in Table 4. In Hyperledger Fabric (HLF), a transition from a Kafka-based ordering service to a Raft-based consensus mechanism was implemented, commencing with version 2.0. This strategic shift aimed to streamline the ordering service and lay the groundwork for the potential future integration of Byzantine Fault Tolerance (BFT) capabilities. While neither Raft nor Kafka inherently possess Byzantine fault tolerance properties, the Raft algorithm's design facilitates the inclusion of such mechanisms in future versions of HLF. This establishes the basis for potentially enhancing the consensus protocol within HLF to incorporate Byzantine fault tolerance, particularly for permissioned blockchain deployments. Our evaluation reveals that the Raft ordering service demonstrably outperformed the Solo approach in terms of performance when processing a diverse range of transaction volumes and rates.

Further considering the impact of programming language on throughput or latency, no substantial variation can be seen since both the programming languages Node.js and Go have their own set of advantages and shortcomings. However, the complexity of chaincode certainly has a great impact on the performance of the HLF blockchain, considering that the complex chaincode may necessitate more computing resources, resulting in slower transaction processing times. The hardware resources available to network nodes might also have an impact on performance. Nodes with faster CPUs, larger memory, and quicker disc I/O can process transactions more quickly and govern more concurrent requests. In our research, we observed that network bandwidth played a significant role in influencing latency. Specifically, the available network bandwidth had a direct impact on the time it took for transactions to propagate and be processed within the HLF network.

In Fig. 14, Case 1 of our study, we examined the impact of operating within a restricted network bandwidth, which resulted in higher latency compared to Case 2, where the system was evaluated under conditions of high network bandwidth. HLF relies on a substantial amount of network bandwidth to ensure efficient operation. This necessity arises from the requirement of distributing transactions to all nodes in the network and validating them.

When the network experiences congestion or has limited bandwidth availability, it can significantly impact the speed of transaction processing and introduce latency issues.

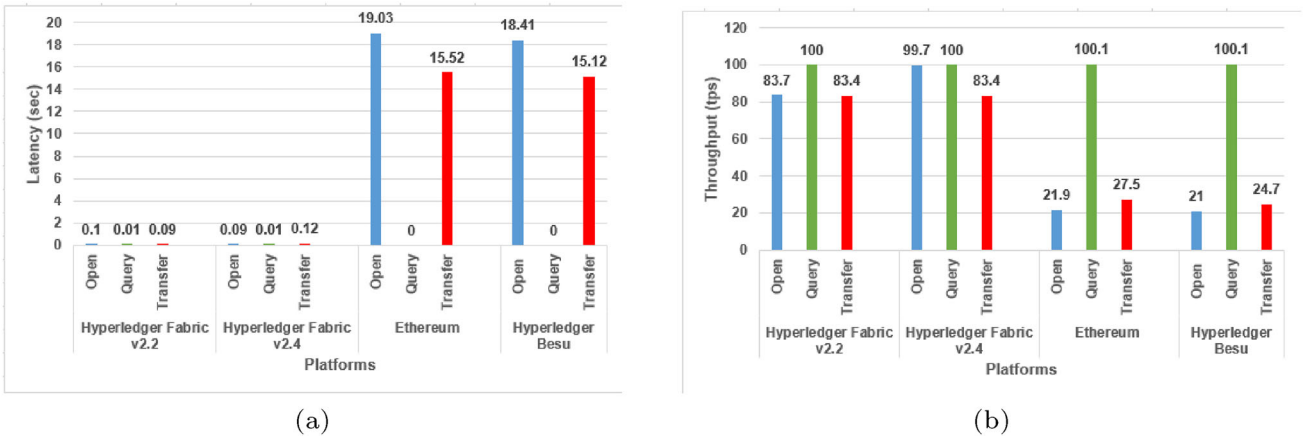


Fig. 13 Latency (average latency) and throughput measured for different platforms for 1000 TXNs transactions [10]

Table 4 Comparison of key features of various consensus mechanisms in ordering service

Features	Solo	Kafka	Raft	PBFT
Byzantine Fault Tolerance	No	No	No	Yes
Crash Fault Tolerance	No	Yes	Yes	Yes
Network Parition Tolerance	No	High	High	Medium
Configuration Complexity	Low	Medium	Low	High
Performance	Limited	High	High	Medium (Increased overhead)
Scalability	Medium	High	High	Medium

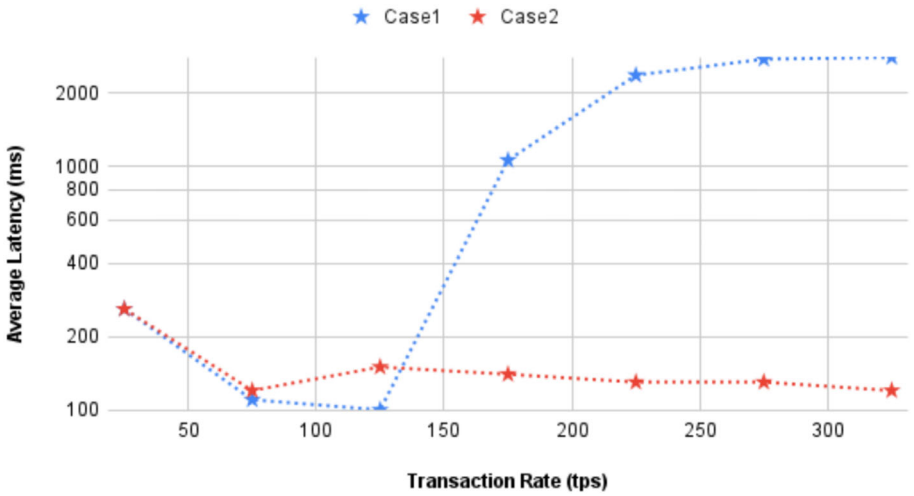


Fig. 14 Impact of network bandwith on latency

6 Conclusion and future work

We conclude that blockchain is gaining exponential popularity among tech savvies. Numerous business applications such as digital identification, secure information exchange, regulatory compliance, asset tokenization, anti-counterfeiting, contract management, and many more are incorporating the blockchain to enhance the product. Choosing the appropriate platform following the need that provides the best performance in all aspects is essential. In our study, we have investigated the impact of workload characteristics on the average latency, resource consumption, and throughput of HLF. The findings suggest that the system works effectively for average latency and throughput between the transaction rate of 150 TPS to 200 TPS and is able to handle 100,000 transactions with minimal delay.

The architecture of HLF is built upon the principles of the Deterministic Consensus Algorithm. This implies that when a new block is sent by the ordering node to any peer for validation, it is ensured to be both final and accurate. This design principle assures the consistency and reliability of the blockchain ledger, as all nodes in the network give consent to the ledger's state. During the evaluation, it was observed that the Raft ordering service outperformed the Solo ordering service, demonstrating higher throughput and lower latency. Our study will help in knowing the effect of several elements on the scalability and performance of HLF version 2.4.9. The HLF is a popular blockchain platform intended for developing enterprise-grade distributed ledger applications. In addition, an investigation of the scalability of the network elucidated that as the number of organizations in the HLF network increased, the throughput decreased and the latency increased. Our study provides statistics on the performance and scalability such as average latency, resource consumption, and throughput. In our investigation, we concluded that several other factors also have a notable influence on the performance of HLF. These factors include CPU power, disk space, the design of chaincode, and network bandwidth. To gain a deeper understanding of performance and scalability in permissioned blockchains, we will expand our future studies. This will involve incorporating factors like network latency, security, and fault tolerance. Additionally, we will evaluate various endorsement policies and consensus protocols under diverse workloads.

Acknowledgements Not Applicable.

Author contributions AT: Implementation, Experimentation, manuscript writing, Writing - Original DRaft, Conceptualization, Methodology, Visualization. VR: Conceptualization, Methodology,

Writing - Review & Editing, Supervision, Investigation. RA: Writing - Review & Editing, Supervision, Investigation.

Funding Not applicable.

Data availability The authors confirm that the data supporting the results and findings of this study are available within the article.

Declarations

Conflict of interest The authors state that they have no known financial or personal interests that could have impacted the conclusions described in this work.

Consent for publication All authors have approved the manuscript and given their full consent for its publication.

Ethical approval and consent to participate There is no research involving human or animal subjects in the article.

References

1. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (2008)
2. Wen, Y., Lu, F., Liu, Y., Huang, X.: Attacks and countermeasures on blockchains: A survey from layering perspective. Computer Networks **191**, 107978 (2021). https://doi.org/10.1016/J.COMNET.2021.107978
3. Bashir, I.: Mastering Blockchain. Packt Publishing Ltd, Birmingham (2017)
4. Bouraga, S.: A taxonomy of blockchain consensus protocols: a survey and classification framework. Expert Syst. Appl. **168**, 114384 (2021). https://doi.org/10.1016/j.eswa.2020.114384
5. Fan, C., Ghaemi, S., Khazaei, H., Musilek, P.: Performance evaluation of blockchain systems: a systematic survey. IEEE Access **8**, 126927–126950 (2020). https://doi.org/10.1109/ACCESS.2020.3006078
6. Sayeed, S., Marco-Gisbert, H.: Assessing blockchain consensus and security mechanisms against the 51% attack. Appl. Sci. **9**(9), 1178 (2019)
7. Crosby, M., Pattanayak, P., Verma, S., Kalyanaraman, V.: Blockchain technology: beyond bitcoin. Appl. Innov. **2**(6–10), 71 (2016)
8. Nguyen, C.T., Hoang, D.T., Nguyen, D.N., Niyato, D., Nguyen, H.T., Dutkiewicz, E.: Proof-of-stake consensus mechanisms for future blockchain networks: fundamentals, applications and opportunities. IEEE Access **7**, 85727–85745 (2019). https://doi.org/10.1109/ACCESS.2019.2925010
9. Vashchuk, O., Shuwar, R.: Pros and cons of consensus algorithm proof of stake. Difference in the network safety in proof of work and proof of stake. Electron. Inf. Technol. **9**(9), 106–112 (2018)
10. Thakur, A., Ranga, D., Agarwal, R.: Performance benchmarking and analysis of blockchain platforms. SSRN Electron. J. (2023). https://doi.org/10.2139/SSRN.4385643
11. Nasir, Q., Qasse, I.A., Abu Talib, M., Nassif, A.B.: Performance analysis of hyperledger fabric platforms. Secur. Commun. Netw. (2018). https://doi.org/10.1155/2018/3976093
12. Thakkar, P., Nathan, S., Viswanathan, B.: Performance benchmarking and optimizing hyperledger fabric blockchain platform. In: 2018 IEEE 26th international symposium on modeling, analysis, and simulation of computer and telecommunication systems (MASCOTS) (2018). https://doi.org/10.1109/MASCOTS.2018.00034
13. Baliga, A., Solanki, N., Verekar, S., Pednekar, A., Kamat, P., Chatterjee, S.: Performance characterization of hyperledger fabric. In: 2018 Crypto Valley Conference on Blockchain Technology (CVCBT), pp. 65–74 (2018). IEEE
14. Sukhwani, H., Wang, N., Trivedi, K.S., Rindos, A.: Performance modeling of hyperledger fabric (permissioned blockchain network). In: 2018 IEEE 17th International Symposium on Network Computing and Applications (NCA) (2018). https://doi.org/10.1109/NCA.2018.8548070
15. Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., Yellick, J.: Hyperledger fabric: a distributed operating system for permissioned blockchains. In: Proceedings of the Thirteenth EuroSys Conference (pp. 1–15) (2018). https://doi.org/10.1145/3190508.3190538.
16. Kuzlu, M., Pipattanasomporn, M., Gurses, L., Rahman, S.: Performance analysis of a hyperledger fabric blockchain framework: throughput, latency and scalability. In: IEEE International Conference on Blockchain (Blockchain) (2019). https://doi.org/10.1109/Blockchain.2019.00003
17. Shalaby, S., Abdellatif, A.A., Al-Ali, A., Mohamed, A., Erbad, A., Guizani, M.: Performance evaluation of hyperledger fabric. In: 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT) (2020). https://doi.org/10.1109/ICIoT48696.2020.9089614
18. Nakaike, T., Zhang, Q., Ueda, Y., Inagaki, T., Ohara, M.: Hyperledger fabric performance characterization and optimization using goleveldb benchmark. In: 2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC) (2020). https://doi.org/10.1109/ICBC48266.2020.9169454
19. Jiang, L., Chang, X., Liu, Y., Mišić, J., Mišić, V.B.: Performance analysis of hyperledger fabric platform: a hierarchical model approach. peer-to-peer networking and applications 13, 1014–1025 (2020) https://doi.org/10.1007/s12083-019-00850-z
20. Foschini, L., Gavagna, A., Martuscelli, R. G.and Montanari: Hyperledger fabric blockchain: Chaincode performance analysis. In: ICC 2020-2020 IEEE International Conference on Communications (ICC) (2020). https://doi.org/10.1109/ICC40277.2020.9149080
21. Pongnumkul, S., Siripanpornchana, C., Thajchayapong, S.: Performance analysis of private blockchain platforms in varying workloads. In: 2017 26th International Conference on Computer Communication and Networks (ICCCN) (2017). https://doi.org/10.1109/ICCCN.2017.8038517
22. Dreyer, J., Fischer, M., Tönjes, R.: Performance analysis of hyperledger fabric 2.0 blockchain platform. In: Proceedings of the Workshop on Cloud Continuum Services for Smart IoT Systems (2020). https://doi.org/10.1145/3417310.3431398
23. Melo, C., Oliveira, J.F.D., Araujo, J., Pereira, P., Maciel, R., Maciel, P.: Performance and availability evaluation of the blockchain platform hyperledger fabric. J. Supercomput. **78**(10), 12505–12527 (2022). https://doi.org/10.1007/s11227-022-04361-2
24. Rouhani, S., Deters, R.: Performance analysis of ethereum transactions in private blockchain. In: 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS) (2017). https://doi.org/10.1109/ICSESS.2017.8342866
25. Huang, D., Ma, X., Zhang, S.: Performance analysis of the raft consensus algorithm for private blockchains. IEEE Trans. Syst. Man Cybernetics Syst. **50**(1), 172–181 (2019). https://doi.org/10.1109/TSMC.2019.2895471
26. Cao, B., Zhang, Z., Feng, D., Zhang, S., Zhang, L., Peng, M., Li, Y.: Performance analysis and comparison of pow, pos and dag based blockchains. Digit. Commun. Netw. **6**(4), 480–485 (2020). https://doi.org/10.1016/j.dcan.2019.12.001
27. Moschou, K., Theodouli, A., Terzi, S., Votis, K., Tzovaras, D., Karamitros, D., Diamantopoulos, S.: Performance Evaluation of

different Hyperledger Sawtooth transaction processors for Blockchain log storage with varying workloads. In: 2020 IEEE International Conference on Blockchain (Blockchain) (2020). https://doi.org/10.1109/Blockchain50366.2020.00069
28. Hang, L., Kim, D.H.: Optimal blockchain network construction methodology based on analysis of configurable components for enhancing hyperledger fabric performance. Blockchain Res. Appl. **2**(1), 100009 (2021). https://doi.org/10.1016/j.bcra.2021.100009
29. Guggenberger, T., Sedlmeir, J., Fridgen, G., Luckow, A.: An in-depth investigation of the performance characteristics of hyperledger fabric. Comput. Ind. Eng. **173**, 108716 (2022). https://doi.org/10.1016/j.cie.2022.108716
30. Wu, O., Li, S., Liu, L., Zhang, H., Zhou, X., Lu, Q.: Performance Modeling of Hyperledger Fabric 2.0. In: Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022 (2022). https://doi.org/10.1145/3530019.3531341
31. Fan, C., Lin, C., Khazaei, H., Musilek, P.: Performance Analysis of Hyperledger Besu in Private Blockchain. In: 2022 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS) (2022). https://doi.org/10.1109/DAPPS55202.2022.00016
32. Wen, Y.F., Hsu, C.M.: A performance evaluation of modular functions and state databases for hyperledger fabric blockchain systems. J. Supercomput. **79**(3), 2654–2690 (2023). https://doi.org/10.1007/s11227-022-04762-3
33. Wang, C., Chu, X.: Performance characterization and bottleneck analysis of hyperledger fabric. In: 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS) (2020). https://doi.org/10.1109/ICDCS47774.2020.00165
34. Hao, Y., Li, Y., Dong, X., Fang, L., Chen, P.: Performance analysis of consensus algorithm in private blockchain. In: 2018 IEEE Intelligent Vehicles Symposium (IV) (2018). https://doi.org/10.1109/IVS.2018.8500557
35. Sondhi, S., Saad, S., Shi, K., Mamun, M., Traore, I.: Evaluating permissioned blockchain using stochastic modeling and chaos engineering. arXiv preprint arXiv:2301.07527https://doi.org/10.48550/arXiv.2301.07527
36. Kaur, M., Khan, M.Z., Gupta, S., Noorwali, A., Chakraborty, C., Pani, S.K.: Mbcp: Performance analysis of large scale mainstream blockchain consensus protocols. IEEE Access **9**, 80931–80944 (2021). https://doi.org/10.1109/ACCESS.2021.3085187
37. Hyperledger: Hyperledger blockchain performance metrics white paper. https://www. hyperledger. org/resources/publications/blockchain-performance-metrics 31(01), 2020 (2018)
38. Singh, N.: Cpu power and network bandwidth-aware optimal block size computation for blockchain-based applications using meta-heuristic algorithms. J. Supercomput. (2023). https://doi.org/10.1007/s11227-023-05210-6
39. Benchmark, C.: https://github.com/hyperledger/caliper-benchmarks. Accessed 26 Mar 2023

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Anita Thakur received her B.Tech. in Computer Science and Engineering from the University Institute of Information Technology and her M.Tech. from Himachal Pradesh University, Summerhill Shimla-05, Himachal Pradesh. She has been enrolled in a Ph.D. program since 2022 and is pursuing her doctorate at Delhi Technological University in Delhi. Her research areas include blockchain, distributed computing, and cryptography.

Virender Ranga received his Ph.D. degree on February 25, 2016, from the Computer Engineering Department of the National Institute of Technology (NIT), Kurukshetra, Under the Ministry of Education, Govt. of India, Haryana, India, followed by M.Tech. and B.Tech. degrees. He has been working as an Associate Professor since October 2021 in the Information Technology Department of Delhi Technological University, Delhi. Previously, he worked as an Assistant Professor with the Computer Engineering Department of the National Institute of Technology Kurukshetra, Haryana, India, from March 20, 2008, to October 25, 2021. He has published more than 100 research articles in various SCI/SCIE/SCOPUS/ESCI/INSPEC Journals and various reputed International Conferences in Computer Communications and Computer Security areas. He is an active reviewer of various reputed journals like IEEE Transaction journals, Springer journals, Elsevier journals, Taylor & Francis journals, Wiley journals, and InderScience journals. His research area includes Wireless Sensor and ad-hoc network security, IoT security, FANET security, SDN security, IoRT, NLP, ITS, etc.

Ritu Agarwal received her Ph.D. in Information Technology in October 2021 from the Information Technology Department at Delhi Technological University. Currently, she is working as an Assistant Professor in the Department of Information Technology at Delhi Technological University, Delhi. Her research interests include Data Structures, Algorithm design, Information Security, Data Recovery, Digital Media Forensics, and Cyber Forensics.

Catalysis Letters
https://doi.org/10.1007/s10562-024-04754-3

Zinc-Copper-Nickel Mixed Metal Oxide as Heterogeneous Catalytic Material for the Reductive Degradation of Nitroarene and Azo Dye

Jigyasa Pathak[1] · **Poonam Singh**[1]

Received: 1 April 2024 / Accepted: 12 June 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

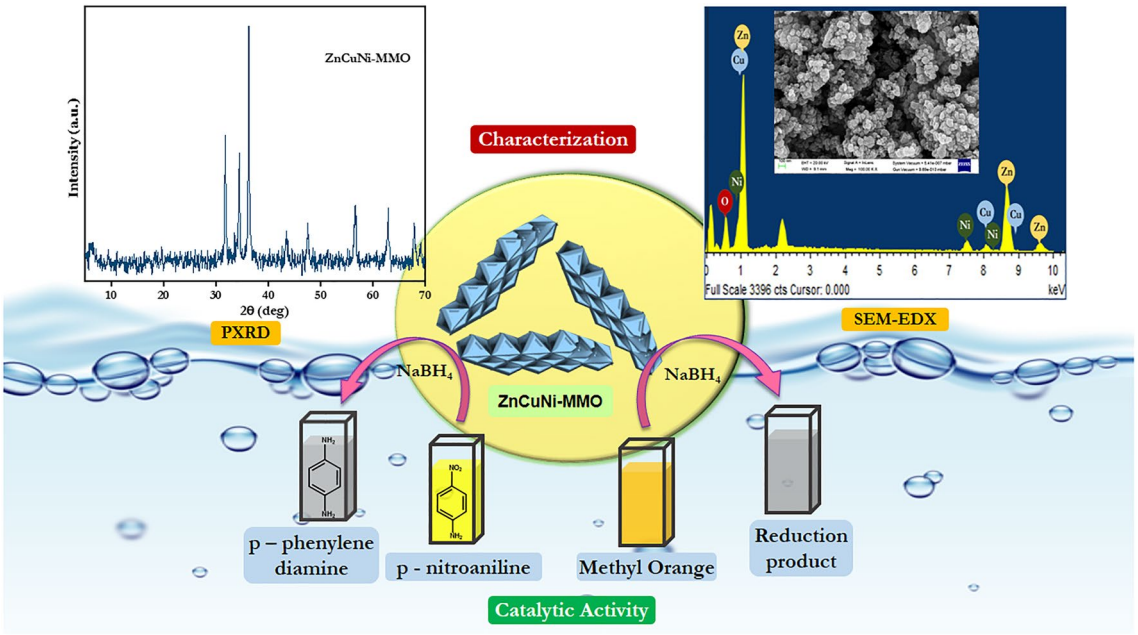
Transition metal-based mixed metal oxides (MMOs) are nexus nanomaterials that garner significant interest from scientists because of their unique magnetic, electronic, optical and catalytic properties that can easily be tailored by varying their composition and structure. Although MMOs hold significant potential in multifunctional applications, but they are plagued by certain challenges such as identifying the appropriate method for synthesis, complications in controlling the surface area and the oxidation states of the constituent transition metals, while also ensuring the homogenous distribution of the constituent metal ions. Therefore, the present work aims to study the formation of homogenous and porous zinc-copper-nickel mixed metal oxide (ZnCuNi-MMO) by performing calcination of ZnCuNi-LDH at 350 °C. The obtained ZnCuNi-MMO was characterized using PXRD, SEM–EDX and BET techniques. Thereafter, ZnCuNi-MMO was applied as a heterogeneous catalyst for the hydrogenation of *p*-nitroaniline (*p*-NA) and catalytic reduction of methyl orange (MO) dye. The pollutant degradation characteristics were assessed using time-dependent UV–Visible absorption spectroscopy showing advanced efficient behavior of ZnCuNi-MMO towards the hydrogenation of *p*-NA (96.98%) and reduction of MO (95.58%). The catalyst exhibited fast reaction rates (0.402 min^{-1} for hydrogenation of *p*-NA and 0.471 min^{-1} for catalytic reduction of MO) and kinetics analysis of the experimental data was found to be coherent with the pseudo-first order model, thereby implying that the catalysis proceeded through the Langmuir–Hinshelwood mechanism. Thus the obtained experimental results highlight the utility and viability of synthesized MMO as an efficacious and sustainable catalytic material.

✉ Poonam Singh
poonam@dtu.ac.in

[1] Department of Applied Chemistry, Delhi Technological University, Delhi 110042, India

Published online: 27 June 2024

Graphical Abstract



Keywords Layered double hydroxide · Mixed metal oxide · Catalytic reduction · *p*-Nitroaniline · Azo dye · Wastewater treatment

1 Introduction

Nitroarene compounds and azo dyes are few of the most pernicious pollutants that currently threaten the health of our ecosystem. Their uninhibited release into water bodies from textile, cosmetics, pharmaceutical, leather, and tanning industries greatly contributes to the deterioration of quality of water bodies. Further, their chemical and thermal inertness, along with the high toxicity of these pollutants necessitates the need for their elimination from water sources. A variety of processes such as adsorption, coagulation-flocculation, filtration, etc. have been developed for their elimination, but they are often found to be ineffective and lead to formation of secondary pollutants, thereby limiting their practical applications [1–4]. Furthermore, drawbacks such as particle agglomeration, membrane fouling, difficulty in flocculant removal, etc. plague traditional treatment methods and call for the development of alternate water remediation strategies [5].

Recently, catalytic degradation of pollutants has emerged as an effective, economical, and green methodology for water remediation since the catalytic reduction of azo dyes yields carbon dioxide and water as benign by-product [6]. While the catalytic treatment of nitroarene compounds such as *p*-nitroaniline yields commercially useful by-products such as *p*-phenylenediamine that can be further utilized for the fabrication of cosmetics, engineering polymers and composites [7–9]. Thus, catalysis can prove to be a beneficial strategy for pollutant sequestration. This methodology often utilizes sodium borohydride ($NaBH_4$) as a reducing agent along with a catalytic material. But on its own, $NaBH_4$ exhibits slow reaction rates for reduction reactions since the reaction is kinetically unfeasible, but in the presence of suitable catalysts, reaction rates are enhanced dramatically. For this purpose, the use of various classes of materials such as metal oxides, noble metals, and carbonaceous materials as catalysts has been previously reported [10–13]. But their commercial applications are quite limited due to the drawbacks such as high cost agglomeration tendency, material instability, and low catalytic efficiency.

In recent years, layered double hydroxides (LDHs) have emerged as new generation heterogeneous catalytic materials since they possess unique structural features and exceptional properties such as high surface area, ion exchange capacity, memory effect property, etc. [14]. They are two-dimensional layered materials consisting of positively charged layers and have anions and water molecules present in the interlayer region. Positive charge on the brucite-like layers of LDHs appears by virtue of the substitution of divalent cations by trivalent cations, thereby creating a charge deficit on the layers. This charge deficit is offset by the interlayer anions that are sandwiched in between the LDH layers along with

water molecules. Their anion and cation exchange capabilities make them extremely versatile and tailorable materials for catalytic applications [15–17]. Furthermore, the incorporation of transition metals in LDHs imparts extraordinary optical, electronic, and magnetic properties to these materials due to the synergistic interactions that arise in the lattice. However, these materials sometimes tend to become less effective as catalysts due to catalytic poisoning by organic moieties.

Therefore, in order to further enhance the tenability and utility of LDHs, an effective methodology is their conversion into mixed metal oxides (MMOs) using thermal treatment methods. MMOs synthesized using conventional methods often possess non-uniform distribution of metal cations, and possess high agglomeration tendencies. Conversely, LDH precursor-based MMOs possess enhanced surface areas and homogeneous distribution of constituent metals, leading to excellent catalytic abilities [18, 19]. Additionally, MMOs containing transition metals also exhibit a myriad of exemplary catalytic properties due to the synergistic interactions between the transition metals in the lattice. Specifically, the addition of zinc, copper and nickel in MMOs can prove to be a boon since they are cost effective and abundant in nature. Apart from this, they offer properties having electron transport and relay abilities, broad band gaps (3.3 – 4.0 eV), high mechanical and thermal stabilities, and electron defects [20–22]. Zn, Cu and Ni play a key role in reduction of organic pollutants since they possess special electronic structures, high surface areas and large number of active sites, due to which they can adsorb various molecules on their surface and activate them. Furthermore, oxidation states of Cu and Ni can be easily tailored, as result of which they are often employed for the fabrication of various homogenous (eg. organometallic complexes) and heterogeneous catalysts (eg. LDHs, Metal–organic frameworks (MOF)) by combining them with a variety of anionic, cationic and neutral ligands. Additionally, their morphologies can be easily tailored, giving rise to special microstructures and they often possess pyro- and piezo-electric properties along with special electronic, optical and magnetic properties that can prove to be beneficial for a wide array of catalytic applications [23–25]. For instance, Murugahandham et al. reported the synthesis of ZnO nanobundles and employed them for the catalytic ozonation of 2-ethyl ethoxy acetate [26]. Belay et al. synthesized zero-valent copper and copper oxides using different precursors and studied their catalytic ability for the reduction of NO_3^- to NO_2^- [27]. In another study, Bhosale et al. reported the use of NiO nanorods for the catalytic synthesis of substituted benzimidazole, benzoxazole and benzothiazole [28]. But inspite of their several favourable characteristics, the combined use of Zn, Cu and Ni in LDHs and LDH derivatives remains limited. Furthermore, although few reports describing the use of LDHs for catalytic reduction of azo dyes and *p*-nitroaniline are available in literature, but virtually no reports describe the use of LDH-derived ZnCuNi-MMOs for catalytic reduction of organic pollutants [29–31].

Therefore, the objectives of the current study include the synthesis and characterization of ZnCuNi-MMO that was synthesized in accordance with our previous report, as presented in Fig. 1 [32]. The synthesized MMO was applied as a heterogeneous catalyst for the hydrogenation of *p*-NA and reduction MO pollutants, and kinetics analysis for the catalysis reactions were carried out. It was observed that ZnCuNi-MMO exhibited excellent removal efficiencies and fast reaction rates for the hydrogenation of *p*-nitroaniline and catalytic reduction of methyl orange in short intervals of time.

2 Experimental

2.1 Synthesis of ZnCuNi-MMO

ZnCuNi-MMO was prepared by the thermal treatment of ZnCuNi-LDH precursor as per the methodology described by Pathak et al. [32]. Briefly, 4.249 mmol zinc acetate dehydrate (Merck, 98.00% purity), 0.37 mmol copper acetate monohydrate (Merck, 98.00% purity) and 0.37 mmol nickel acetate tetrahydrate (CDH chemicals, 98.00% purity) were separately dissolved in 10 mL distilled water and then mixed with each other with constant stirring. Then 5 mL of 30% H_2O_2 and 30 mL distilled water were added to the above solution and it was sonicated for 30 min. Afterwards, the reaction mixture was stirred for 24 h and subsequently, it was dried completely at 60 °C. The obtained ZnCuNi-LDH was then calcined in muffle furnace at 350 °C for 6 h to obtain black-greyish colored ZnCuNi-MMO.

2.2 Catalytic Activity of ZnCuNi-MMO

The catalytic prowess of synthesized ZnCuNi-MMO was evaluated by studying the hydrogenation of *p*-nitroaniline (*p*-NA) and reduction of methyl orange (MO) dye as model reactions in presence of sodium borohydride ($NaBH_4$).

For the hydrogenation of *p*-nitroaniline (*p*-NA), *p*-NA solution (1 mL, 1 mM) was mixed well with 1 mL of distilled water followed by the addition of freshly prepared $NaBH_4$ solution (1 mL, 100 mM). After this, the catalytic reaction was initiated by addition of with 0.001 g ZnCuNi-MMO to the reaction mixture. Progress of hydrogenation reaction was determined by recording the absorption spectrum of *p*-NA at $\lambda_{max} = 380$ nm using UV–visible spectrophotometer at 1 min intervals till solution was completely decolorized.

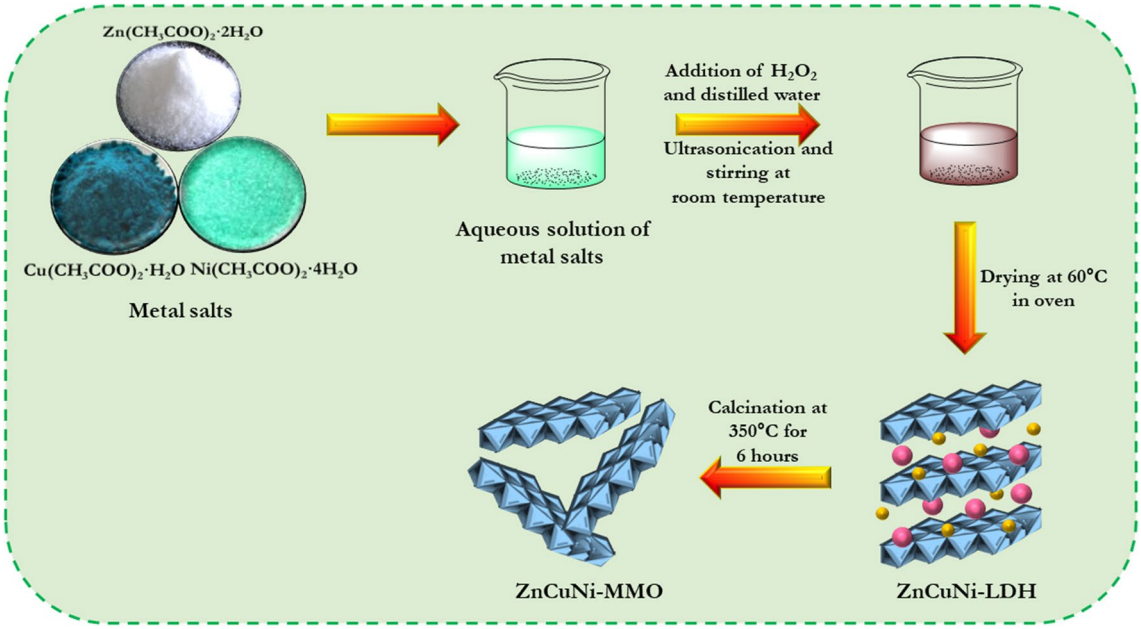


Fig. 1 Schematic representation of synthesis of ZnCuNi-MMO

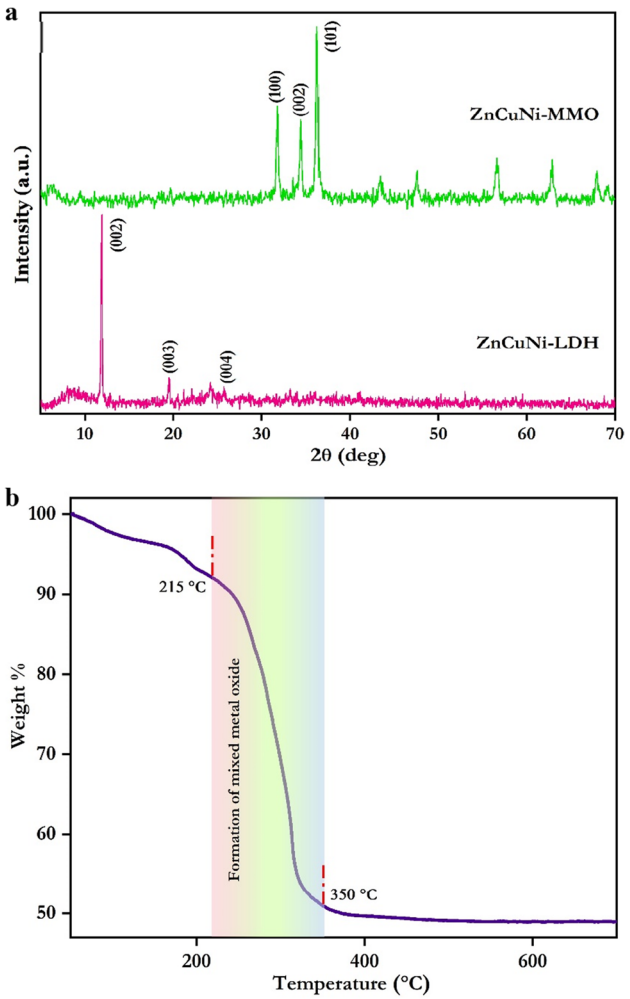


Fig. 2 **a** Comparitive plot of PXRD patterns of ZnCuNi-LDH and ZnCuNi-MMO **b** TGA thermogram of ZnCuNi-LDH

The catalytic reduction of MO dye using ZnCuNi-MMO was evaluated in a similar manner. MO solution (1 mL, 100 ppm) was mixed well with 1 mL distilled water and 1 mL of freshly prepared 100 mM $NaBH_4$ solution. Then 0.001 g MMO was added to reaction mixture and reaction progress was determined by recording absorption spectrum at $\lambda_{max} = 466$ nm at 1 min intervals till complete decolorization of solution.

The percentage degradation of *p*-NA and MO was calculated using the formula:

$$Percentage/degradation = \frac{(C_O - C_f)}{C_O} \times 100 \quad (1)$$

where C_O and C_f denote the initial and final concentrations of *p*-NA and MO.

3 Results and Discussion

3.1 Characterisation Results

The PXRD diffractogram of ZnCuNi-LDH presented in Fig. 2a was found to be consistent with the typical hydrotalcite pattern exhibiting sharp intense peaks at low *2θ* angle and low intensity peaks at higher *2θ* angle, corresponding to *(002), (003)* and *(004)* hkl planes. The appearance of these peaks corroborated the formation of highly crystalline LDH in pure phase. The basal spacing distance for *(002)* plane was found to be 7.41 Å, that was coherent with previously

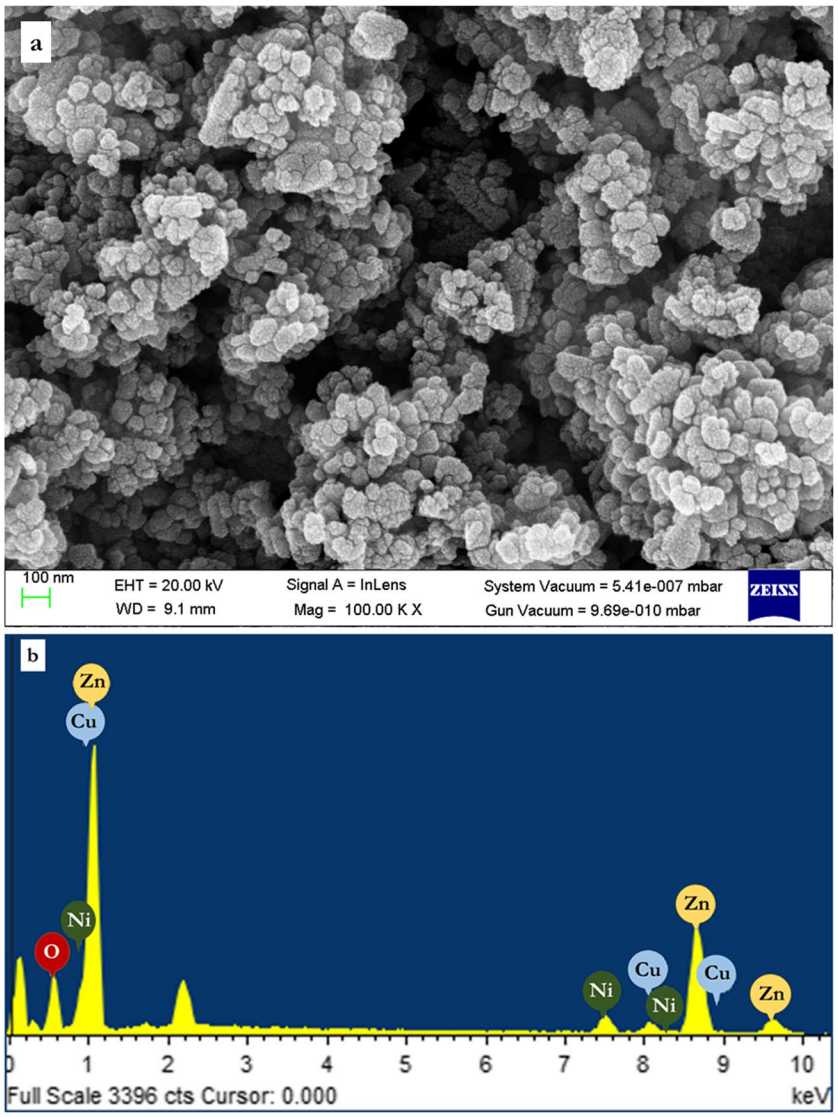


Fig. 3 **a** SEM micrograph and **b** EDX spectrum of ZnCuNi-MMO

reported values in literature [33]. The thermal stability of ZnCuNi-LDH precursor was ascertained from the TGA thermogram presented in Fig. 2b. Thermal degradation of ZnCuNi-LDH included the initial loss of physisorbed and interlayer water molecules upto 215 °C, followed by the decarboxylation, dehydroxylation and removal of interlayer acetate ions from the LDH lattice upto 350 °C leading to the breakdown of LDH layers and formation of pure phase mixed metal oxide (MMO) [34]. Based on TGA results obtained for the LDH, 350 °C was chosen as the optimal temperature for calcination of ZnCuNi-LDH to synthesize ZnCuNi-MMO.

Upon calcination, collapse of layered structure was observed, leading to the disappearance of characteristic peaks of LDH in PXRD pattern (Fig. 2a). The emergence of new peaks at 31.8°, 34.56° and 36.39° corresponding to *(100), (002)* and *(101)* hkl planes of hexagonal wurtzite structure of crystalline ZnO with space group $P6_3mc$ was observed. Thus, the PXRD results support the formation of pure and homogenous ZnCuNi-MMO upon the complete decomposition of LDH on thermal treatment [35]. The average crystallite size was calculated using the Debye–Scherrer equation and was found to be 24.89 nm.

The surface morphology and microstructure of ZnCuNi-MMO was examined using SEM technique. The SEM micrograph presented in Fig. 3a shows the absence of the plate-liked morphology responsible for layered structure of LDH and clearly indicates presence of irregular spherical aggregates, thereby confirming the formation of mixed metal oxide (MMO) [36]. Further, the presence of Zn, Cu and Ni was confirmed using EDX that clearly showed their homogeneous distribution in the lattice (Fig. 3b).

Surface area is a critical parameter that monitors the catalytic activity of a material. The specific surface area (S_{BET}) and pore structure of ZnCuNi-MMO were

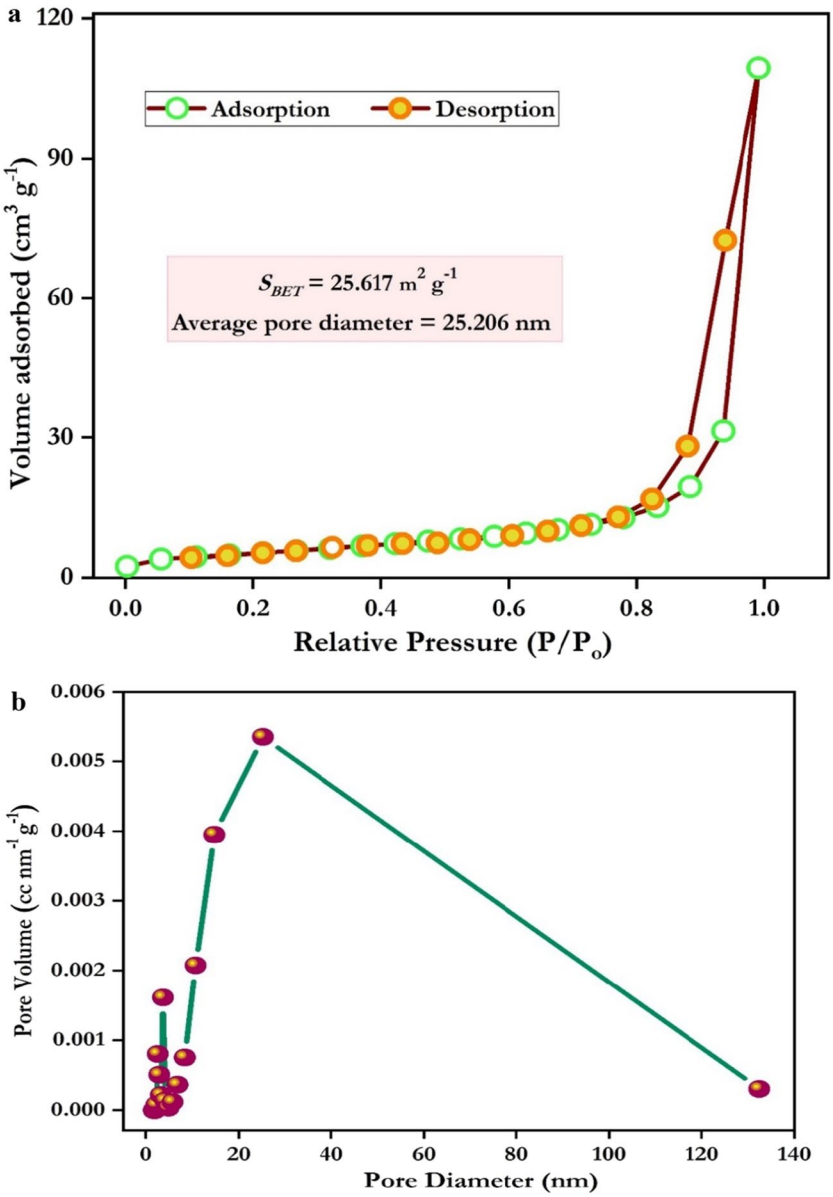


Fig. 4 **a** N_2 adsorption–desorption isotherm and **b** corresponding pore size distribution plots of ZnCuNi-MMO

investigated using N_2 adsorption–desorption study by BET analysis (Fig. 4a-b). S_{BET} was found to be 25.617 m^2g^{-1}, while the values of pore volume and average pore diameter, calculated using the BJH method, were found to be 0.174 cc g^{-1} and 25.206 nm respectively. The shape of the isotherm is similar to Type-IV, as described by IUPAC classification, and it was inferred that synthesized ZnCuNi-MMO possessed mesoporosity. The appearance of H3-type hysteresis loop indicated the existence of narrow slit-shaped pores in the mesoporous ZnCuNi-MMO aggregates [37]. The formation of pores may be a consequence of the calcination of ZnCuNi-LDH. Therefore, the textural properties described above render ZnCuNi-MMO as suitable materials for catalytic applications.

3.2 Catalytic Activity

The catalytic prowess of ZnCuNi-MMO for the $NaBH_4$-assisted hydrogenation of *p*-NA and reduction of MO was examined using time dependent UV–Visible spectrophotometry. In the catalytic experiment, the rate of reaction is assumed independent of $NaBH_4$ concentration since $NaBH_4$ (100 mM) was added in excess. Therefore, kinetics of the hydrogenation of *p*-NA and catalytic reduction of MO were calculated in accordance with the pseudo-first order model presented in Eq. (2) as follows:

$$\ln\left(\frac{C_f}{C_O}\right) = K \tag{2}$$

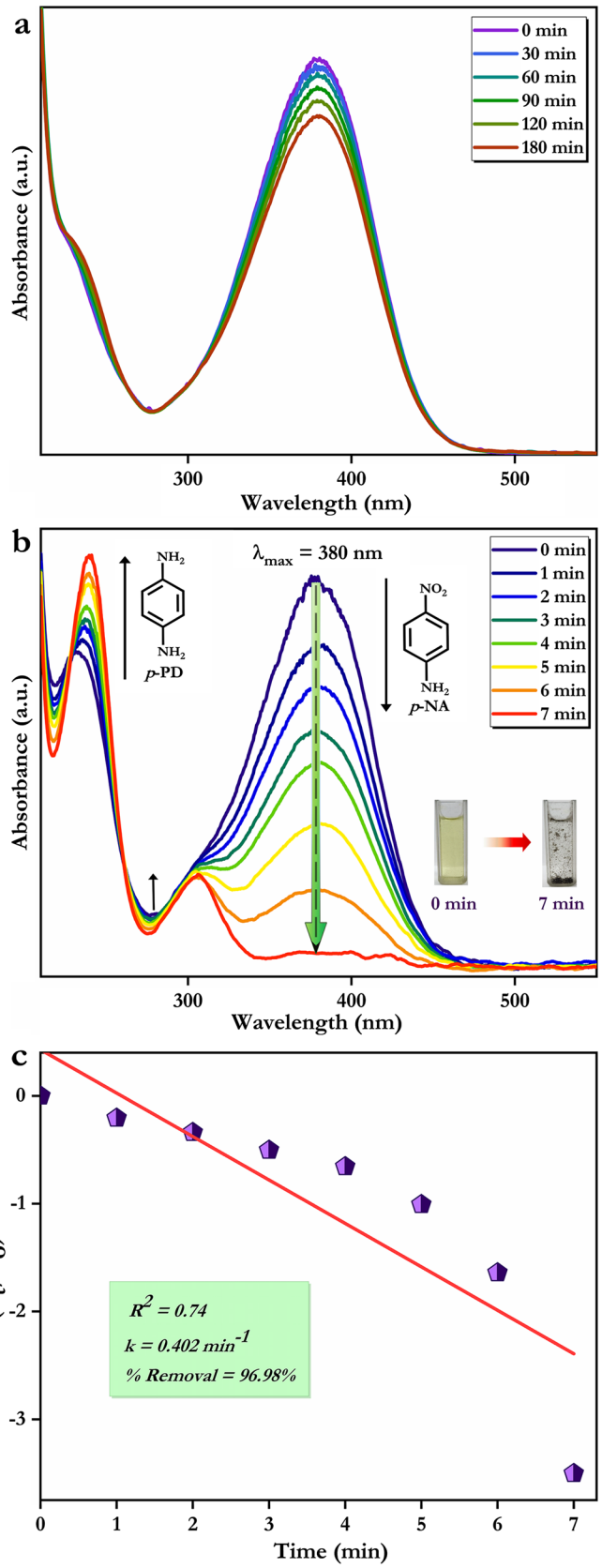
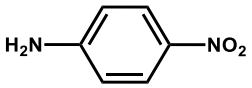
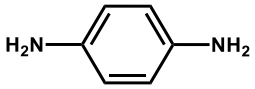
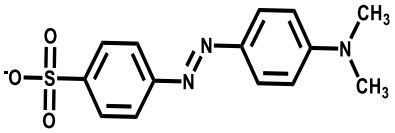
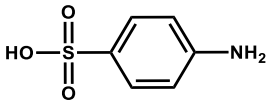


Fig. 5 **a** Time-dependent UV–Vis absorption spectrum of hydrogenation of *p*-NA using $NaBH_4$, **b** Time-dependent UV–Vis absorption spectrum and **c** linear regression of pseudo-first kinetics model for the $NaBH_4$ – assisted hydrogenation of *p*-NA using ZnCuNi-MMO

Table 1 Structures of *p*-NA and MO with the potential products formed on their catalytic treatment with $NaBH_4$ and ZnCuNi-MMO

Organic Pollutant	λ_{max} of pollutant (nm)	By-product	λ_{max} of by-product (nm)
p-Nitroaniline 	380	*p*-phenylenediamine 	240 and 304
Methyl orange 	466	Sulfanilic Acid 	250

where C_O and C_f are the initial and final concentrations of *p*-NA and MO, while K denotes rate constant.

Since the reaction solutions possess distinct colours, their reaction rate can be evaluated in terms of comparative absorptive intensity based on the Beer-Lambert law (Eq. (3))

$$A = \varepsilon Cl \tag{3}$$

$$\ln\left(\frac{C_f}{C_O}\right) = \ln\left(\frac{A_f}{A_O}\right) = kt \tag{4}$$

where A_O and A_f are the initial and final absorbance of *p*-NA and azo dyes.

3.2.1 Catalytic Hydrogenation of *p*-NA

NACs such as *p*-NA possess high toxicity and chemical inertness, thereby making them extremely harmful to plant and animal life [38]. Thus, it is extremely necessary to eliminate these pollutants from water bodies. Also, the hydrogenation product of *p*-NA i.e. *p*-phenylenediamine (*p*-PD) is an important commercial product [39–41]. Therefore, the catalytic hydrogenation of *p*-NA can serve a two-fold purpose. Thus, the hydrogenation of *p*-NA by ZnCuNi-MMO to form *p*-phenylenediamine was evaluated in the presence of excess $NaBH_4$. The time dependent UV–Visible spectrum was recorded at $\lambda_{max} = 380$ nm to monitor the hydrogenation of *p*-NA.

Initially, hydrogenation of *p*-NA was examined in the presence of $NaBH_4$ in order to determine the role of catalyst. The time dependent UV–Visible spectra of *p*-NA hydrogenation using $NaBH_4$ is presented in Fig. 5a. It was observed that when only $NaBH_4$ was added to reaction mixture, there was a very slow decrease in the absorbance of *p*-NA at $\lambda_{max} = 380$ nm leading to only 13.86% hydrogenation of *p*-NA even after 3 h. However, upon addition of ZnCuNi-MMO to the reaction mixture, it was observed that the intensity of peak at $\lambda_{max} = 380$ nm corresponding to *p*-NA began decreasing at a rapid rate and eventually disappeared after 7 min (Fig. 5b). The percentage removal efficiency, calculated using Eq. (1) was found to be 96.98%. The hydrogenation of *p*-NA due to conversion of the nitro group to amine group ($-NH_2$) was further confirmed by visual examination of the reaction mixture that showed change in colour from bright yellow to colourless. Further, appearance of new peaks at 240 nm and 304 nm confirmed the formation of *p*-phenylenediamine (*p*-PD) that is the hydrogenation product of *p*-NA (Table 1) [42]. Therefore, it was concluded that addition of ZnCuNi-MMO to the reaction mixture was essential for the efficient hydrogenation of *p*-NA. The rate of reaction was evaluated by monitoring the decrease in absorbance of *p*-NA at $\lambda_{max} = 380$ nm, and rate constant of reduction (k) was determined by linearly plotting ln (C_f/C_o) versus reaction time t (min). Therefore, k was calculated from the kinetics data using Eq. (4) by employing the pseudo first order model and was found to be 0.402 min^{-1} ($R^2 = 0.74$) (Fig. 5c).

3.2.2 Methyl Orange Reduction

Methyl orange, an anionic azo dye, is a common pH indicator and textile dyeing agent. The presence of azo (-N=N-) chromophore group is responsible for the bright orange

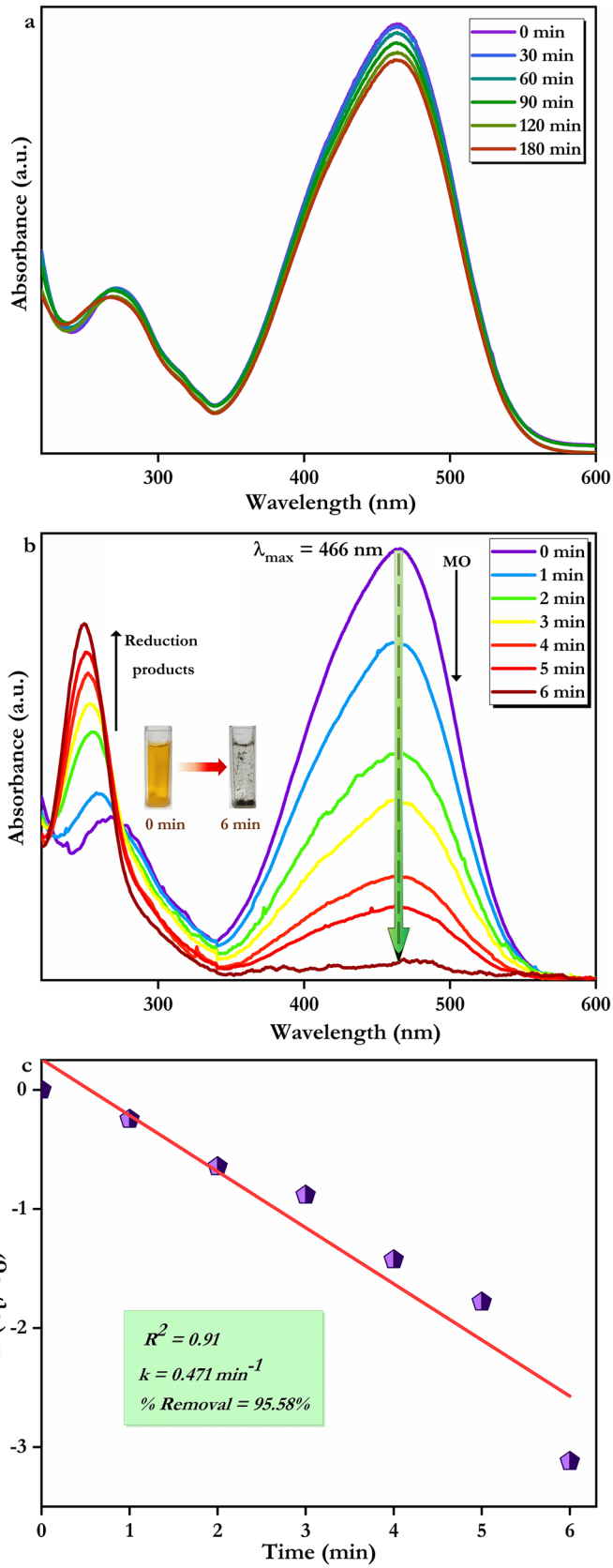


Fig. 6 **a** Time-dependent UV–Vis absorption spectrum of catalytic reduction of MO using $NaBH_4$, **b** Time-dependent UV–Vis absorption spectrum and **c** linear regression of pseudo-first kinetics model for the $NaBH_4$ – assisted catalytic reduction of MO using ZnCuNi-MMO

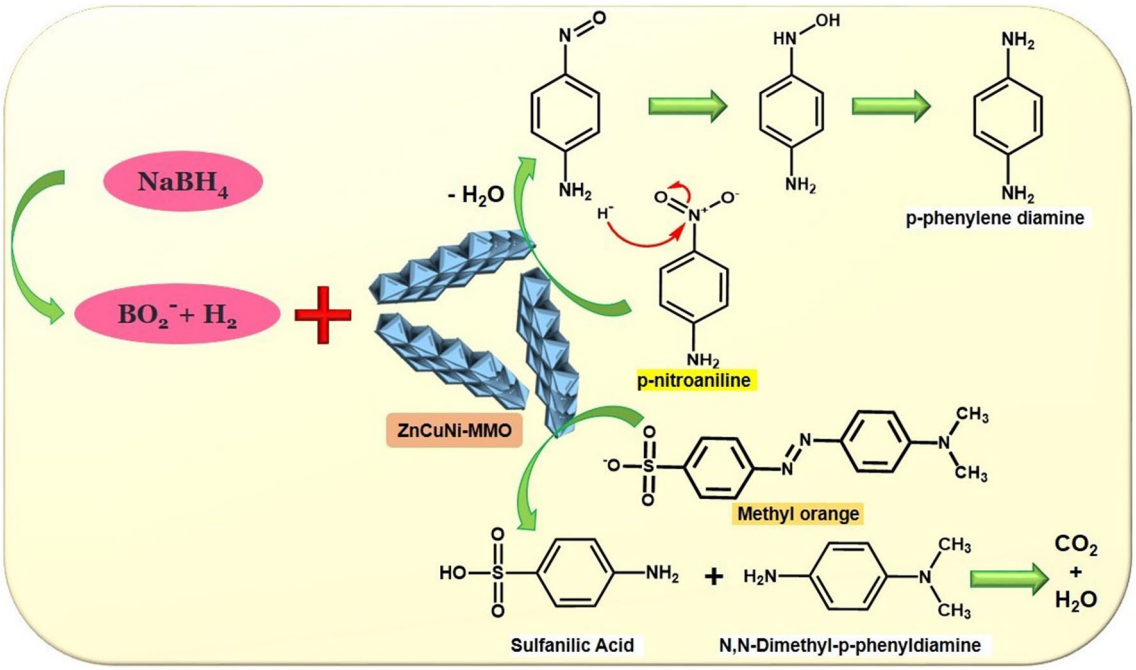


Fig. 7 Postulated mechanism for the $NaBH_4$- assisted catalytic degradation of *p*-NA and MO using ZnCuNi-MMO

color of MO dye. According to several reports, MO has been found to be toxic and excessive exposure to MO may cause skin and eye irritation, nausea, and diarrhea [43, 44]. Hence, the elimination of MO from water sources has become an issue of much concern.

Initially, catalytic reduction of MO was examined in the presence of $NaBH_4$ and time dependent UV–Visible spectrum was recorded at $\lambda_{max}=466$ nm to determine the extent of MO reduction by ZnCuNi-MMO (Fig. 6a). It was observed that addition of only $NaBH_4$ to reaction mixture leads to a very slow decrease in the absorbance of MO at $\lambda_{max}=466$ nm leading to only 8.76% reduction of MO even after 3 h. Subsequently, on addition of ZnCuNi-MMO to the reaction mixture. With addition of ZnCuNi-MMO catalyst, the intensity of peak corresponding to MO dye present at $\lambda_{max}=466$ nm began to decrease rapidly until its disappearance after 6 min corresponding to the complete decolorisation of the reaction mixture. Percentage reduction of MO, calculated using Eq. (1), was found to be 95.58%. The formation of by-products was corroborated by the emergence of a new peak at 250 nm and its intensity kept increasing with time (Fig. 6b). Therefore, it was concluded that addition of ZnCuNi-MMO to the reaction mixture was essential for the efficient reduction of MO. The rate constant *k*, calculated from Eq. (4) in accordance with pseudo first order kinetics, was evaluated to be 0.471 min^{-1} ($R^2=0.91$) (Fig. 6c). It may be worth noting that the interaction of $NaBH_4$ with MO molecules results in only partial reduction of the azo group, while the simultaneous action of $NaBH_4$ and ZnCuNi-MMO catalyst is successful in breaking the chromophore azo (-N=N-) group of MO and leads to the formation of smaller amine moieties. The appearance of absorbance peaks around 250 nm may be attributed to the possible formation of hydrazine derivatives as reduction products, potentially including sulfanilic acid (Table 1) [12]. The disappearance of peak at 466 nm and the emergence of new peak at 250 nm corroborates the reduction of MO by the catalytic action of ZnCuNi-MMO.

3.2.3 Postulated Mechanism

The catalytic experiments show that ZnCuNi-MMO can act as an excellent catalyst for the $NaBH_4$ assisted hydrogenation of *p*-NA and catalytic reduction of MO. The catalysis process is known to follow the Langmuir–Hinshelwood mechanism, which includes the adsorption of reactions on catalyst surface followed by the interaction between the adsorbed reactants to form products. The potential mechanism for the $NaBH_4$- assisted hydrogenation of *p*-NA and catalytic reduction of MO is presented in Fig. 7.

$NaBH_4$ is a good reducing agent and on addition to water, $NaBH_4$ produces nucleophilic borohydride (BH_4^-) anions. These BH_4^- anions breakdown into BO_2^- that deliver electrons to the catalyst and H_2 that get adsorbed on the MMO surface. H_2 adsorbed on ZnCuNi-MMO surface forms a metal-hydride complex that aids in the transfer of H_2 and electrons to the $–NO_2$ group. This leads to the hydrogenation of $–NO_2$ group to $–NH_2$, and results in the formation of *p*-phenylenediamine.

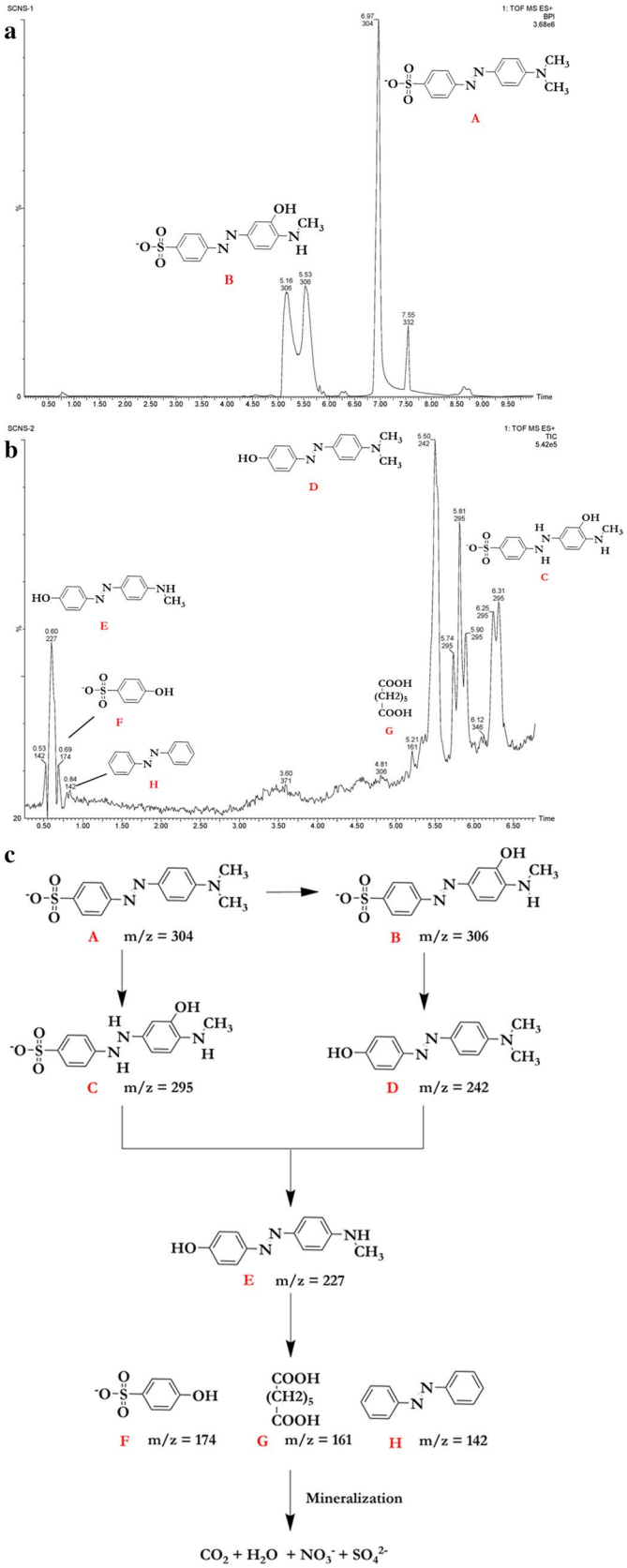


Fig. 8 LC–MS chromatograms of **a** untreated MO (0 min) and **b** by-products obtained on $NaBH_4$- assisted catalytic reduction of MO using ZnCuNi-MMO and **c** potential degradation pathway of MO

Table 2 Comparison of removal efficiencies of ZnCuNi-MMO with previously reported catalysts

Pollutant	Catalyst	Time (min)	Removal Efficiency	Reference
p-NA	Bi_2S_3	10	96%	[48]
	Au nanowire networks	17	Complete reduction	[49]
	RGO-Ni Nanocomposite	190	93%	[50]
	ZnCuNi-MMO	7	96.98%	This work
MO	Asp-Au NPs	10	> 95%	[12]
	NiFe2O4/γ-Fe2O3	36	Complete reduction	[51]
	Ag–Cu Rhizome Powder	9	96%	[52]
	ZnCuNi-MMO	6	95.58%	This work

While in case of MO, the dye molecules are adsorbed onto the ZnCuNi-MMO surface. ZnCuNi-MMO acts as an electron donor while the dye molecules act as an electron acceptor and get reduced to its hydrazine derivatives such as sulfanilic acid. Finally, the reduction products are desorbed from MMO surface, thereby regenerating it for further catalytic activity.

The products formed on the catalytic reduction of *p*-NA and MO are benign, can be easily degraded, less toxic, do not pollute, and hold commercial importance. Therefore, catalytic reduction of *p*-NA and MO using ZnCuNi-MMO in presence of $NaBH_4$ is a favourable methodology for the elimination of these pollutants from aqueous medium.

3.3 Degradation Products of Methyl Orange

The degradation products of MO dye were examined using LC–MS to gain a better understanding of the degradation process. The LC–MS chromatograms of untreated MO (0 min) and the reaction mixture after complete decolourisation of MO (6 min) in ESI + mode are presented in Fig. 8. MO ($C_{14}H_{14}N_3NaO_3S$) gets converted to $C_{14}H_{14}N_3SO_3^-$ due to the dissociation of Na^+ in water. In the LC–MS chromatogram presented in Fig. 8a, the parent molecule MO exhibits a clear signal at $m/z = 304$ (A). Additionally, the signal at $m/z = 306$ (B) may correspond to the presence of monohydroxylated MO. Further, the LC–MS chromatogram presented in Fig. 8b clearly shows the appearance of signals corresponding to by-products of MO degradation. The signals appearing at $m/z = 295$ (C), 242 (D), 227 (E), 174 (F), 161 (G) and 147 (H) correspond to the by-products (illustrated in Fig. 8c) that are formed as a result of the degradation of MO in the presence of $NaBH_4$ and ZnCuNi-MMO [45–47].

Based on the potential degradation pathway of MO presented in Fig. 8c, it may be assumed that the degradation of MO consists of processes including bond cleavage, demethylation, hydroxyl group insertion, detachment of sulfonic group, benzene ring opening and finally the mineralization of intermediates to potentially form CO_2, H_2O, NO_3^-, SO_4^{2-}, etc.

3.4 Comparison With Reported Catalysts

The comparison of the catalytic activity of ZnCuNi-MMO with previously reported catalysts towards the hydrogenation of *p*-NA and catalytic reduction of MO is presented in Table 2. The rate constants for reactions catalyzed by ZnCuNi-MMO were found to be comparable or even better than the rate constants of most reported catalysts. Therefore, it may be concluded that ZnCuNi-MMO holds the potential to be an excellent candidate for the catalytic reduction of nitroarene compounds and azo dyes. The superior catalytic activity may be attributed to the small size, high surface area and homogeneous distribution of Zn, Cu and Ni in ZnCuNi-MMO catalyst.

4 Conclusion

In summary, the current study reports the facile synthesis of homogeneous MMO having uniform distribution of Zn, Cu, and Ni by thermal treatment of ZnCuNi-LDH precursor. The high specific surface area of 25.617 m^2g^{-1} estimated using BET indicates that they may hold practical importance for catalytic applications. Keeping in view the above observations, ZnCuNi-MMO was applied as a catalyst for carrying out the hydrogenation of *p*-NA and catalytic reduction of MO dye. ZnCuNi-MMO was able to successfully hydrogenate *p*-NA to *p*-phenylenediamine, and reduce MO to its hydrazine derivatives including sulfanilic acid respectively with efficiency of more than 95% within ~ 7 min. The reaction rates, evaluated using pseudo-first order model, were found to be comparatively higher than already existing catalysts i.e. 0.402 min^{-1} (*p*-NA hydrogenation) and 0.471 min^{-1} (MO dye reduction). The catalytic action of ZnCuNi-MMO with the intermediates BO_2^- and H_2 that are produced on the hydrolysis of $NaBH_4$ enable its action as an electron relay and facilitate the formation of metal-hydride complexes that catalyze the hydrogenation of *p*-NA and reduction of MO. Further, LC–MS technique was employed to identify potential products formed on the catalytic reduction of MO, that are generally considered to be benign and less toxic, and

the degradation pathway for MO was elucidated. Therefore, these favourable attributes of ZnCuNi-MMO indicate their multifarious applicability for the degradation of nitroarene compounds and organic dyes in water remediation, and they may be considered as potential candidates for other catalytic applications.

Acknowledgements The authors would like to express their immense gratitude to Delhi Technological University for providing the resources and infrastructure to conduct this research work. Author JP would like to thank Delhi Technological University for providing fellowship.

Authors Contributions Conceptualization [Jigyasa Pathak] and [Poonam Singh], Methodology Development [Jigyasa Pathak], Funding Acquisition [Poonam Singh], Sample and Data Analysis [Jigyasa Pathak], Supervision and Validation [Poonam Singh], Writing—original draft [Jigyasa Pathak], Writing—review & editing [Jigyasa Pathak] and [Poonam Singh].

Funding Authors are grateful to UGC BSR Startup Grant No. F.30–416/2018 (BSR) for the funds to carry out this research. Author JP would like to thank Delhi Technological University for providing fellowship.

Data Availability Data available upon request.

Declarations

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent to Publish Not applicable.

Competing Interests The authors have no financial or proprietary interests in any material discussed in this article.

References

1. Ahmad A, Mohd-Setapar SH, Chuong CS et al (2015) Recent advances in new generation dye removal technologies: novel search for approaches to reprocess wastewater. RSC Adv 5:30801–30818. https://doi.org/10.1039/c4ra16959j
2. Ma B, Xue W, Hu C et al (2019) Characteristics of microplastic removal via coagulation and ultrafiltration during drinking water treatment. Chem Eng J 359:159–167. https://doi.org/10.1016/j.cej.2018.11.155
3. Pathak J, Singh P (2022) Adsorptive removal of Congo red using organically modified zinc–copper–nickel ternary metal hydroxide: kinetics, isotherms and adsorption studies. J Polym Environ 31:327–344. https://doi.org/10.1007/s10924-022-02612-0
4. dos Santos GE de S, Lins PV dos S, Oliveira LMT de M et al (2021) Layered double hydroxides/biochar composites as adsorbents for water remediation applications: recent trends and perspectives. J Clean Prod 284:124755. https://doi.org/10.1016/j.jclepro.2020.124755
5. Khan SA, Shah LA (2021) Advanced removal techniques for dye-containing wastewaters. Springer Nature Singapore, Gateway
6. Takehira K (2017) Recent development of layered double hydroxide-derived catalysts – rehydration, reconstitution, and supporting, aiming at commercial application –. Appl Clay Sci 136:112–141. https://doi.org/10.1016/j.clay.2016.11.012
7. Lu X, Li L, Song B et al (2015) Mechanistic investigation of the graphene functionalization using p-phenylenediamine and its application for supercapacitors. Nano Energy 17:160–170. https://doi.org/10.1016/j.nanoen.2015.08.011
8. Sk MM, Yue CY (2014) Layer-by-layer (LBL) assembly of graphene with p-phenylenediamine (PPD) spacer for high performance supercapacitor applications. RSC Adv 4:19908–19915. https://doi.org/10.1039/c4ra02652g
9. Wang JJ, Jiang J, Hu B, Yu SH (2008) Uniformly shaped poly(p-phenylenediamine) microparticles: shape-controlled synthesis and their potential application for the removal of lead ions from water. Adv Funct Mater 18:1105–1111. https://doi.org/10.1002/adfm.200700583
10. Wang Y, Gao P, Wei Y et al (2021) Silver nanoparticles decorated magnetic polymer composites (Fe3O4@PS@Ag) as highly efficient reusable catalyst for the degradation of 4-nitrophenol and organic dyes. J Environ Manag 278:111473. https://doi.org/10.1016/j.jenvman.2020.111473
11. Ahmad I, Aalam G, Amir M et al (2022) Development of highly efficient magnetically recyclable Cu2+/Cu0 nano-photocatalyst and its enhanced catalytic performance for the degradation of organic contaminations. Sci Total Environ 846:157154. https://doi.org/10.1016/j.scitotenv.2022.157154
12. Garg N, Bera S, Rastogi L et al (2020) Synthesis and characterization of L-asparagine stabilised gold nanoparticles: catalyst for degradation of organic dyes. Spectrochim Acta - Part A Mol Biomol Spectrosc 232:118126. https://doi.org/10.1016/j.saa.2020.118126
13. Sun L, Xiang X, Wu J et al (2019) Bi-metal phosphide NiCoP: An enhanced catalyst for the reduction of 4-nitrophenol. Nanomaterials 9.https://doi.org/10.3390/nano9010112
14. Fan G, Li F, Evans DG, Duan X (2014) Catalytic applications of layered double hydroxides: recent advances and perspectives. Chem Soc Rev 43:7040–7066. https://doi.org/10.1039/c4cs00160e
15. Tang S, Yao Y, Chen T et al (2020) Recent advances in the application of layered double hydroxides in analytical chemistry: a review. Anal Chim Acta 1103:32–48. https://doi.org/10.1016/j.aca.2019.12.065
16. Pathak J, Singh P (2024) Ultrafast catalytic reduction of organic pollutants using ternary zinc–copper–nickel layered double hydroxide. Appl Organomet Chem 1–17. https://doi.org/10.1002/aoc.7507
17. Silva Neto LD, Anchieta CG, Duarte JLS et al (2021) Effect of drying on the fabrication of MgAl layered double hydroxides. ACS Omega 6:21819–21829. https://doi.org/10.1021/acsomega.1c03581
18. Zhang S, Liu Q, Fan G, Li F (2012) Highly-dispersed copper-based catalysts from Cu-Zn-Al layered double hydroxide precursor for gas-phase hydrogenation of dimethyl oxalate to ethylene glycol. Catal Lett 142:1121–1127. https://doi.org/10.1007/s10562-012-0871-8
19. Gabriel R, Carvalho SHV d, Duarte JL da S et al (2022) Mixed metal oxides derived from layered double hydroxide as catalysts for biodiesel production. ApplCatal A Gen 630:118470. https://doi.org/10.1016/j.apcata.2021.118470
20. Vinitha V, Preeyanghaa M, Vinesh V et al (2021) Two is better than one: catalytic, sensing and optical applications of doped zinc oxide nanostructures. Emergent Mater 4:1093–1124. https://doi.org/10.1007/s42247-021-00262-x
21. Narender SS, Varma VVS, Srikar CS et al (2022) Nickel oxide nanoparticles: a brief review of their synthesis, characterization, and applications. Chem Eng Technol 45:397–409. https://doi.org/10.1002/ceat.202100442

22. Scotti N, Monticelli D, Zaccheria F (2012) Dispersed copper oxide: a multifaceted tool in catalysis. Inorganica Chim Acta 380:194–200. https://doi.org/10.1016/j.ica.2011.10.001
23. Sun Y, Chen L, Bao Y et al (2016) The applications of morphology controlled ZnO in catalysis. Catalysts 6:188. https://doi.org/10.3390/catal6120188
24. Ye RP, Lin L, Li Q et al (2018) Recent progress in improving the stability of copper-based catalysts for hydrogenation of carbon-oxygen bonds. Catal Sci Technol 8:3428–3449. https://doi.org/10.1039/c8cy00608c
25. Ma Y, Ma C, Wang Y, Wang K (2022) Advanced nickel-based catalysts for urea oxidation reaction: challenges and developments. Catalysts 12.https://doi.org/10.3390/catal12030337
26. Muruganandham M, Wu JJ (2008) Synthesis, characterization and catalytic activity of easily recyclable zinc oxide nanobundles. Appl Catal B Environ 80:32–41. https://doi.org/10.1016/j.apcatb.2007.11.006
27. Belay TA, Lin CY, Hsiao HM et al (2018) Effect of precursor type on the reduction of concentrated nitrate using zero-valent copper and sodium borohydride. Water Sci Technol 77:114–122. https://doi.org/10.2166/wst.2017.525
28. Bhosale MA, Bhanage BM (2015) Rapid synthesis of nickel oxide nanorods and its applications in catalysis. Adv Powder Technol 26:422–427. https://doi.org/10.1016/j.apt.2014.11.015
29. Leandro SR, Marques IJ, Torres RS et al (2021) Nitroarene and dye reduction with 2:1 Co/Al layered double hydroxide catalysts – Is gold still necessary? Inorg Chim Acta 521:120336. https://doi.org/10.1016/j.ica.2021.120336
30. Nam HJ, Park EB, Jung DY (2016) Bioinspired polydopamine-layered double hydroxide nanocomposites: controlled synthesis and multifunctional performance. RSC Adv 6:24952–24958. https://doi.org/10.1039/c5ra28103b
31. Wei Z, Li Y, Dou L et al (2019) Cu3- xNixAl-layered double hydroxide-reduced graphene oxide nanosheet array for the reduction of 4-nitrophenol. ACS Appl Nano Mater 2:2383–2396. https://doi.org/10.1021/acsanm.9b00273
32. Pathak J, Singh P (2022) Synthesis and characterization of ternary layered double hydroxide containing zinc/copper/nickel and its PANI composite. Polym Compos 43:7836–7844. https://doi.org/10.1002/pc.26895
33. Singh P, Nagarajan R (2015) Synthesis and characterization of hydrotalcite type structure containing Zn2+ and La3+-ions. Mater Lett 159:58–60. https://doi.org/10.1016/j.matlet.2015.06.071
34. Kazeem TS, Zubair M, Daud M et al (2019) Graphene/ternary layered double hydroxide composites: efficient removal of anionic dye from aqueous phase. Korean J Chem Eng 36:1057–1068. https://doi.org/10.1007/s11814-019-0284-0
35. Suresh S, Thambidurai S, Arumugam J et al (2022) Antibacterial activity and photocatalytic oxidative performance of zinc oxide nanorods biosynthesized using Aerva lanata leaf extract. InorgChem Commun 139. https://doi.org/10.1016/j.inoche.2022.109398
36. Annapoorani A, Koodalingam A, Beulaja M et al (2022) Eco-friendly synthesis of zinc oxide nanoparticles using Rivina humilis leaf extract and their biomedical applications. Process Biochem 112:192–202. https://doi.org/10.1016/j.procbio.2021.11.022
37. Li Z, Zhang Q, Liu X et al (2018) Mechanochemical synthesis of novel heterostructured Bi2S3/Zn-Al layered double hydroxide nano-particles as efficient visible light reactive Z-scheme photocatalysts. Appl Surf Sci 452:123–133. https://doi.org/10.1016/j.apsusc.2018.04.237
38. Gu Y, Wang Y, Zhang H (2018) Study on the interactions between toxic nitroanilines and lysozyme by spectroscopic approaches and molecular modeling. Spectrochim Acta - Part A Mol Biomol Spectrosc 202:260–268. https://doi.org/10.1016/j.saa.2018.05.008
39. Ahmed HAM, Maaboud RMA, Latif FFA et al (2013) Different analytical methods of para-phenylenediamine based hair dye. J Cosmet Dermatological Sci Appl 03:17–25. https://doi.org/10.4236/jcdsa.2013.33a1003
40. Jaidev RS (2012) Poly(p-phenylenediamine)/graphene nanocomposites for supercapacitor applications. J Mater Chem 22:18775–18783. https://doi.org/10.1039/c2jm33627h
41. Zhang J, Zhang X, Zheng Y (2021) Synthesis of poly(p-phenylenediamine) encapsulated graphene and its application in steel protection. Prog Org Coatings 158:106330. https://doi.org/10.1016/j.porgcoat.2021.106330
42. Jiji SG, Gopchandran KG (2019) Shape dependent catalytic activity of unsupported gold nanostructures for the fast reduction of 4-nitroaniline. Colloids Interface Sci Commun 29:9–16. https://doi.org/10.1016/j.colcom.2018.12.003
43. Aljuaid A, Almehmadi M, Alsaiari AA et al (2023) g-C3N4 based photocatalyst for the efficient photodegradation of toxic methyl orange dye: recent modifications and future perspectives. Molecules 28:3199. https://doi.org/10.3390/molecules28073199
44. Farhan Hanafi M, Sapawe N (2020) A review on the water problem associate with organic pollutants derived from phenol, methyl orange, and remazol brilliant blue dyes. Mater Today Proc 31:A141–A150. https://doi.org/10.1016/j.matpr.2021.01.258
45. An Q, Tang M, Deng S et al (2023) Methyl orange degradation with peroxydisulfate activated with the synergistic effect of the acid-modified red mud and biochar catalyst. Arab J Sci Eng 48:8819–8834. https://doi.org/10.1007/s13369-022-07398-w
46. Ariyanti D, Maillot M, Gao W (2018) Photo-assisted degradation of dyes in a binary system using TiO2 under simulated solar radiation. J Environ Chem Eng 6:539–548. https://doi.org/10.1016/j.jece.2017.12.031
47. Wang B, Xu M, Chi C et al (2017) Degradation of methyl orange using dielectric barrier discharge water falling film reactor. J Adv Oxid Technol 20:20170021
48. Guo F, Ni Y, Ma Y et al (2014) Flowerlike Bi2S3 microspheres: Facile synthesis and application in the catalytic reduction of 4-nitroaniline. New J Chem 38:5324–5330. https://doi.org/10.1039/c4nj00900b
49. Chirea M, Freitas A, Vasile BS et al (2011) Gold nanowire networks: Synthesis, characterization, and catalytic activity. Langmuir 27:3906–3913. https://doi.org/10.1021/la104092b
50. Das P, Ghosh S, Baskey M (2019) Heterogeneous catalytic reduction of 4-nitroaniline by RGO-Ni nanocomposite for water resource management. J Mater Sci Mater Electron 30:19731–19737. https://doi.org/10.1007/s10854-019-02323-8
51. Salaheldin HI (2018) Chitosan molecular weight effects on the synthesis of gold nanoparticles and catalytic degradation of environmental pollutants. J Adv Phys 14:5388–5405
52. Bakhsh EM, Ismail M, Sharafat U et al (2022) Highly efficient and recoverable Ag-Cu bimetallic catalyst supported on taro-rhizome powder applied for nitroarenes and dyes reduction. J Mater Res Technol 18:769–787. https://doi.org/10.1016/j.jmrt.2022.02.062

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.